MASSACHUSETTS
INSTITUTE of TECHNOLOGY
LIBRARIES

HB31
.M415
19-05
28
DEWEY

working paper
department
of economics

CHOOSING THE NUMBER OF INSTRUMENTS

Stephen G. Donald
Whitney K. Newey

No. 99-05 February, 1999

massachusetts
institute of
technology
50 memorial drive
cambridge, mass. 02139

WORKING PAPER
DEPARTMENT
OF ECONOMICS

CHOOSING THE NUMBER OF INSTRUMENTS

Stephen G. Donald
Whitney K. Newey

No. 99-05 February, 1999

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

50 MEMORIAL DRIVE
CAMBRIDGE, MASS. 02142

Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

http://www.archive.org/details/choosingnumberof00dona

1 Introduction

There has been a resurgence of interest in the application and properties of instrumental variables (IV) estimators. An important problem in IV estimation is choosing the number of valid instruments, or more generally a subset of all the instruments known to be valid. The properties of estimators are sensitive to this choice (e.g. see Morimune, 1983), even in applied work with many observations (e.g. see Bound, Jaeger and Baker, 1996). We address this problem by providing simple approximate mean-square error (MSE) criteria, that can be minimized to choose the instruments. We give these criteria for two-stage least squares (2SLS), limited information maximum likelihood (LIML), and the Jacknife IV (JIVE) estimator of Angrist, Imbens, and Kreuger (1995). We also compare the approximate MSE of these estimators, and find LIML is best.

Our criteria is based on higher-order asymptotic theory, like that of Nagar (1959), Anderson and Sawa (1979), Morimune (1983), and Rothenberg (1983). Our approximate MSE criteria are like those of Nagar (1959), being based on the MSE of leading terms in an expansion of the estimator.This approach is well known to give the same answer as the MSE of leading terms in an Edgeworth expansion, under suitable regularity conditions (e.g. Rothenberg, 1984), and has been used in nonparametric and semiparametric models by Andrews (1991a), Linton (1995), and Powell and Stoker (1996). For k-class estimators our calculations extend those of Nagar (1959) and Rothenberg (1983) to the misspecified first-stage case. We also provide new results for the jacknife IV estimator (JIVE) of Angrist, Imbens and Krueger (1995).

A number of recent studies have considered the properties of IV estimators when instruments are only weakly correlated with the endogenous right hand side variables. In particular Nelson and Startz (1990), Maddala and Jeong (1992), Bekker (1994), Bound, Jaeger and Baker (1996), and Staiger and Stock (1997) have shown that standard IV estimators, including the 2SLS estimator, tend to be biased towards the inconsistent OLS estimator and that inferences based on 2SLS can be quite misleading, when the endogenous variable is only weakly correlated with the instrument. As a practical mat-

ter, choosing the instruments may help improve the approximation, particularly when the source of weak correlation is extraneous instruments that should be excluded. For example, we show that in data like that of Angrist and Krueger (1991) our criteria will choose the smallest number of instruments, where there is less evidence of a problem of weak correlation.

Our criteria for choosing a subset of valid instruments is the estimated MSE of leading terms in an expansion. This approach is different than Andrews (1996), who bases instrument choice on the GMM criterion function. We are choosing instruments from a subset that is known to be valid while he is searching for the largest set of valid instruments. Our approach seems ideally suited to many applications in microeconomic data, where there is a large set of instruments all thought to be valid. Examples include the draft lottery number as an instrument for military service, as in Angrist (1990), and interactions of covariates with instruments, as in Angrist and Krueger (1991). Our results also apply to the choice of nonlinear functions to use in the efficient semiparametric instrumental variables estimator of Newey (1990). Here we derive the optimal, MSE minimizing number of instruments to use, answering the important question of how to pick the number of instruments in optimal semiparametric estimation. The number of instruments can be thought of as a smoothing parameter for the nonparametric component, as has been considered by Linton (1995) and Powell and Stoker (1996) for other models.

In Section 2 we describe the IV estimators we consider, present the criteria for instrument choice, and compare these criteria for different estimators, seeking the one with smallest higher order MSE. In section 3 we derive the approximate MSE for LIML, 2SLS and JIVE when there are no covariates. Section 4 studies the properties of the criteria theoretically and shows that they can be implemented in a way that is optimal in a certain sense. Section 5 allows for covariates in the theoretical results. In section 6 we present a Monte-Carlo experiment. Section 7 presents the results of applying the criteria in the Angrist and Krueger (1991) application.

2 The Model, Estimators, and Instrument Selection Criteria

The model that we consider is,

$$\begin{aligned} y_i &= \gamma Y_i + x_{1i}'\beta + \epsilon_i, E[\epsilon_i|x_i] = 0, Var(\epsilon_i|x_i) = \sigma_\epsilon^2, \\ Y_i &= f(x_i) + u_i, E[u_i|x_i] = 0, Var(u_i|x_i) = \sigma_u^2, \end{aligned}$$

for $i = 1, ..., N$, where y_i, and Y_i are scalars, x_i is a $d \times 1$ vector of exogenous variables, and x_{1i} is a $d_1 \times 1$ vector of exogenous variables which are assumed to be a subset of x_i, where we assume homoskedasticity throughout. We also assume that conditional third moments are zero throughout the paper. The first equation is the equation of interest and the right hand side variable Y_i is possibly correlated with ϵ_i so that generally $E(u_i\epsilon_i|x_i) = \sigma_{u\epsilon} \neq 0$, The second equation represents the reduced form relationship between Y_i and the exogenous variables x_i which is allowed to be nonparametric, with $f(x_i) = E(Y_i|x_i)$.

Because of the conditional moment restriction $E(\epsilon_i|x_i) = 0$, functions of x_i can be used as instruments in estimating the equation of interest. Let $\psi_i^K = (\psi_{1K}(x_i), ..., \psi_{KK}(x_i))'$ be a vector of K functions to be used as instrumental variables, where we assume throughout that ψ_i^K includes x_{1i}. They could be approximating functions, such as power series or regression splines, as in Newey (1990). The problem we consider is how to choose this instrument vector so that the associated IV estimators have good properties. For simplicity we have allowed K to serve as both the number of instruments and the index of the instruments, but we could proceed more generally by specifying a different index for the instruments. If we did that the criteria could be used to compare instrumental variables estimators with the same number of instruments.

We consider several different IV estimators. To describe them let $\Psi^K = [\psi_1^K, ..., \psi_n^K]'$ be the matrix of observations on the instrumental variables, $P^K = \Psi^K(\Psi^{K\prime}\Psi^K)^{-1}\Psi^{K\prime}$ the associated projection matrix, $y = (y_1, ..., y_n)'$, $Y = (Y_1, ..., Y_n)'$, $X_1 = [x_{11}, ..., x_{1n}]'$, $W = [Y, X_1]$, and $\delta = (\gamma, \beta')'$. Also, let θ be the minimum of $\lambda'Y'(P^K - P_1)Y\lambda/\lambda'Y'(I - P_1)Y\lambda$

over λ, where $P_1 = X_1(X_1'X_1)^{-1}X_1'$. The main class of estimators considered is the k-class which includes estimators which have the form,

$$\hat{\delta} = ((1+\kappa)W'P^KW - \kappa W'W)^{-1}((1+\kappa)W'P^Ky - \kappa W'y)$$

where the scalar κ is given by,

$$\kappa = \frac{a\theta + \frac{b}{N}}{1 - a\theta - \frac{b}{N}}$$

for constants a and b. This class of estimators includes 2SLS, where $a = b = 0$, LIML, where $a = 1$ and $b = 0$, and a bias-corrected 2SLS estimator like that of Nagar (1959), denoted here by B2SLS, where $a = 0$ and $b = K - (2 + d_1)$. This class of estimators is similar to that considered by Rothenberg (1983).

The other estimator considered is the Jackknife IV estimator (JIVE) proposed by Angrist, Imbens and Krueger (1995). It has the form,

$$\begin{pmatrix} \hat{\gamma}_J \\ \hat{\beta}_J \end{pmatrix} = \begin{pmatrix} Y'C'Y & Y'C'X_1 \\ X_1'Y & X_1'X_1 \end{pmatrix}^{-1} \begin{pmatrix} Y'C'y \\ X_1'y \end{pmatrix}$$

where C is the matrix with $C_{ij} = P_{ij}/(1 - P_{ii})$ for $i \neq j$ and $C_{ii} = 0$. Note that CY has the interpretation as the vector of predictions of the Y_i given all but the ith observation, hence the Jackknife terminology.

The instrument selection criteria are based on the approximate mean square error (MSE) for the endogenous variable coefficient estimator $\hat{\gamma}$. Implementing these criteria requires preliminary estimates of some of the parameters of the model and a goodness of fit criteria for estimation of the reduced form using the instruments ψ_i^K. Let $\hat{\delta}$ denote a preliminary IV estimator and let $\hat{\epsilon} = y - W\hat{\delta}$. Also, let $\hat{u}$ denote some preliminary reduced form residual estimator, such as that obtained from the residuals from the regression of Y on $\Psi^{\bar{K}}$ for some *fixed* $\bar{K}$. Then let

$$\hat{\sigma}_\epsilon^2 = \hat{\epsilon}'\hat{\epsilon}/N, \hat{\sigma}_u^2 = \hat{u}'\hat{u}/N, \hat{\sigma}_{u\epsilon} = \hat{u}'\hat{\epsilon}/N.$$

It is important in what follows that none of these preliminary estimators depend on K. For example $\hat{\delta}$ might be an IV estimator with just one instrument for Y_i, or it might be an IV estimator where the instruments are chosen to minimize the first stage goodness

of fit criteria given below. Similarly, the reduced form residual $\hat{u}$ might be from the first stage regression with the best fit. In any case these estimates must remain fixed as the criteria for different instruments sets is calculated.

The reduced form goodness of fit criteria can be formed in at least two ways. Let $\hat{u}^K$ denote the reduced form residual vector from regressing Y on Ψ^K. The Mallows goodness of fit criterion for the reduced form is

$$\hat{R}(K) = (\hat{u}^{K\prime}\hat{u}^K/N)(1 + \frac{2K}{N})$$

The cross-validation goodness of fit measure is

$$\hat{R}(K) = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{u}_i^K)^2}{(1 - P_{ii}^K)^2}$$

Either of these will have suitable theoretical properties for use in the instrument selection criteria.

The preliminary estimates of covariance parameters can be combined with the reduced form goodness of fit measures to form criteria for the choice of instruments for each of 2SLS, LIML, and JIVE as

$$\begin{aligned}
\hat{S}_{2SLS}(K) &= \hat{\sigma}_{u\epsilon}^2\frac{K^2}{N} + \hat{\sigma}_\epsilon^2(\hat{R}(K) - \hat{\sigma}_u^2\frac{K}{N}) \\
\hat{S}_{LIML}(K) &= \hat{\sigma}_\epsilon^2\left(\hat{R}(K) - \frac{\hat{\sigma}_{u\epsilon}^2}{\hat{\sigma}_\epsilon^2}\frac{K}{N}\right) \\
\hat{S}_{JIVE}(K) &= \hat{\sigma}_\epsilon^2\left(\hat{R}(K) + \frac{\hat{\sigma}_{u\epsilon}^2}{\hat{\sigma}_\epsilon^2}\frac{K}{N}\right)
\end{aligned}$$

For each estimator, choosing K to minimize the corresponding $\hat{S}(K)$ will result in $\hat{\gamma}$ that has relatively small MSE asymptotically. These criteria also will apply when K is not the sole index of the instruments, e.g. for comparing two different sets of instruments that have the same number of instruments. Choosing the estimator where the expression on the right-hand side is smallest will result in the best asymptotic MSE in that case as well.

It is interesting to compare the size of the criteria for different estimators, because this parallels the MSE comparison of the estimator. As both N and K increase, the

LIML and JIVE criteria are both of smaller order than the 2SLS criteria. This reflects that the bias of 2SLS increases with the number of instruments at a higher rate than the other estimators, a result previously established by Morimune (1983). Consequently, for large numbers of instruments both LIML and JIVE should dominate 2SLS in terms of MSE. Also, the criteria for LIML is smaller than that for JIVE, so that LIML is best among these three estimators. As previously shown by Rothenberg (1983) for the fixed instrument case, it turns out that LIML is best median unbiased to the order we consider in the k-class of estimators.

3 The Mean Square Error

For much of this section we focus on the simplest model which is a special case of the model considered in Section 2. In particular we focus on the model:

$$\begin{aligned} y_i &= \gamma Y_i + \epsilon_i \\ Y_i &= f(x_i) + u_i \end{aligned}$$

for $i = 1, ..., N$, where y_i, Y_i and ϵ_i are all scalars and where Y_i is possibly correlated with ϵ_i. In Section 3.5 below we show how the MSE criteria can be extended to models with covariates as considered in Section 2. In this case, the estimator for γ is given by,

$$\hat{\gamma} = ((1+\kappa)Y'PY - \kappa Y'Y)^{-1}((1+\kappa)Y'Py - \kappa Y'y)$$

where,

$$\kappa = \frac{a\theta + \frac{b}{N}}{1 - a\theta - \frac{b}{N}}$$

with θ being the minimum value of, $\lambda'Y'PY\lambda/\lambda'Y'Y\lambda$. The other specific estimator considered is JIVE, which is denoted by,

$$\hat{\gamma}_J = (Y'C'Y)^{-1}Y'C'y$$

where C was described in Section 2. Unfortunately JIVE does not appear to fit into the k-class of estimators so that MSE calculations will have to be performed separately.

3.1 Calculation

Our approach to finding the approximate MSE is similar to that of Nagar (1959). First we normalize the estimator (let $\hat{\gamma}$ denote a generic estimator),

$$\sqrt{N}(\hat{\gamma} - \gamma) = \hat{H}^{-1}\hat{h}$$

where $\hat{H}$ and $\hat{h}$ are different for different estimators and have population counterparts,

$$h = \frac{1}{N} f'\epsilon$$

and

$$H = \frac{f'f}{N}$$

Then we use the expansion,

$$\begin{aligned} \hat{H}^{-1}\hat{h} &= [(I - (H - \hat{H})H^{-1})H]^{-1}(h + (\hat{h} - h)) \\ &= H^{-1}(\hat{h} + (H - \hat{H})H^{-1}h + (H - \hat{H})H^{-1}(\hat{h} - h) + (H - \hat{H})^2H^{-2}h + ... \end{aligned}$$

and note that since H^{-1} is a common factor for all terms in the expansion, we can calculate an approximate MSE by taking expectations of the square of the expression,

$$\begin{aligned} H\sqrt{N}(\hat{\gamma} - \gamma) &= \hat{h} + (H - \hat{H})H^{-1}h \\ &\quad + (H - \hat{H})H^{-1}(\hat{h} - h) + (H - \hat{H})^2H^{-2}h + ... \end{aligned}$$

Then following Nagar (1959) we find the MSE of this expression using the largest (in probability) terms in the square of this expression, although since we are interested in the selection of K and allow $K \to \infty$ we include only those leading terms in the MSE that are pertinent to the choice of K for the estimator of interest. Thus terms that do not depend on K will be omitted. Because of this fact, terms beyond the second will not (in general) contribute to the approximate MSE criteria to be used for selection of K, even though they do contribute to the MSE in the case where one has a fixed K and one obtains an approximation to the MSE to $O(1/N)$. For the purposes of completeness and to enable other types of comparisons we also provide approximate $O(1/N)$ MSE calculations for

fixed K. These calculations extend the work of Nagar (1959) and Rothenberg (1983) to the case where there is a possible misspecification in the first stage. In such a calculation the third and fourth terms do contribute to the approximate (to $O(1/N)$) MSE.[1] Note that the leading term (which is $O(1)$) in the approximate MSE will be

$$E(hh') = \sigma_\epsilon^2 \frac{f'f}{N}$$

and that this will be common to all of our approximate MSE's. Since this term has no bearing on the choice of K and since it is common to all of our estimators we omit the term from the expressions we present and focus on the remaining terms. All calculations are done conditional on the exogenous variables x_i.

The assumptions which are used to derive an approximate MSE are now stated and discussed.

Assumption 1 *Assume that* $\{x_i, u_i, \epsilon_i\}$ *are iid, and satisfy,*

(i) $E(u_i|x_i) = E(\epsilon_i|x_i) = 0$,

(ii) $0 < E(u_i^2|x_i) = \sigma_u^2 < \infty$ *and* $0 < E(\epsilon_i^2|x_i) = \sigma_\epsilon^2 < \infty$,

(iii) $E({u_i}^j {\epsilon_i}^k|x_i) = 0$, *for* $j+$ k *being a positive odd integer such that* $k + j \leq 5$,

(iv) $E(|u_i|^j|\epsilon_i|^k|x_i) < \infty$, *for any positive integers* j *and* k *such that* $k + j \leq 5$,

(v) x_i *have support* $X \subset R^d$ *that is the Cartesian product of compact intervals.*

Assumption 1(i) was noted in Section 2. Assumption 1(ii) is a homoskedasticity assumption that simplifies the calculations while Assumption 1(iii) is a symmetry assumption which implies that both residuals have symmetric distributions. This will be satisfied if the residuals are jointly normally distributed. Assumption 1(iv) requires the existence of at least five moments for the residuals. Assumption 1(v) is standard in the literature on series based nonparametric regression. The next Assumption concerns the nature of the optimal instrument function f.

[1] We thank Tom Rothenberg for this observation.

Assumption 2 *The function* $f(x_i) = E(Y_i|x_i)$ *is such that,*

(i) $f : X \to R$ *and has an extension to all of* R^d *that is* $s > 0$ *times continuously differentiable in* x,

(ii) with probability 1, $0 < c < N^{-1}\sum_{i=1}^{N} f(x_i)^2 < c^{-1}$ *for some small constant* c *uniformly in* N.

Assumption 2(i) is a smoothness condition that allows one to obtain approximation rates for series based methods. The second part of this assumption is an identification assumption that requires that $f(x_i)$ not be identically zero which would result in γ being unidentified. This will be modified when we consider the introduction of covariates into the structural equation.

Assumption 3: *Letting* P_{ii} *denote the* i*th diagonal element of* P, *assume that*

$$\sup_{i \leq N} P_{ii} \to 0$$

as $N \to \infty$ *with probability one.*

It should be stressed that this assumption is not required for one to show that $\sqrt{N}(\hat{\gamma} - \gamma)$ is asymptotically normal, and is only necessary to obtain tractable expressions for the MSE. It requires some restrictions on the rate at which K increases with the sample size that are much stronger than those needed to show asymptotic normality of the estimators. It should be noted however that as long as s in Assumption 3 is sufficiently large the restrictions will allow for K to increase at a rate that results in the fastest possible rate of convergence of the MSE expression to zero.[2]

Our first result gives an approximation to the MSE for a subclass of the k-class estimators we are considering. In obtaining all of our approximations we have restricted attention to k-class estimators that have values of a and b that satisfy certain properties. In particular it is assumed that $a = O(1)$ while $b = O(K)$, which includes all of the

[2]Donald (1997) has shown that for power series one requires that $K^5/N \to 0$ when x_i has a continuous distribution with density that is bounded and bounded away from 0 on X.

estimators mentioned above and ensures that the estimators will be consistent.[3] The first subclass of estimators consists of those for which the approximate bias (to order $N^{-1/2}$) is increasing in K. Unless otherwise stated all MSE will refer to the terms in the large sample large K MSE that are dominant amongst those that increase with K and those that decrease with K.

Proposition 1 *Given Assumptions 1-3, assuming that* $a = O(1)$ *and* $b = O(K)$ *such that* $(1-a)-b/K$ *is bounded away from* 0, *the MSE for* $\sqrt{N}(\hat{\gamma}-\gamma)$ *is approximately*

$$\sigma_{u\epsilon}^2 \frac{[(1-a)K-b]^2}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}.$$

It is straightforward to show that the approximate bias, to $O(N^{-1/2})$ is of the form[4],

$$\frac{[(1-a)K-b-(2-a)]\sigma_{u\epsilon}}{\sqrt{N}} \left(\frac{f'f}{N}\right)^{-1} \tag{1}$$

and that for estimators in the class under consideration whose bias term grows with K, the leading term in the MSE comes directly from the square of this bias. The second term in our approximate MSE comes from the approximation error inherent in the estimation of f by a series based approximation. The result in Proposition 1 leads directly to an approximate MSE for 2SLS which is the only estimator under consideration which falls into this class.

Corollary 1 *For 2SLS,* $\hat{\gamma}_2$, *which has* $a = 0$ *and* $b = 0$, *the approximate MSE of* $\sqrt{N}(\hat{\gamma}_2 - \gamma)$, *under Assumptions 1-3, is given by,*

$$\sigma_{u\epsilon}^2 \frac{K^2}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}.$$

[3]Note that, as will be seen, these conditions on their own do not ensure root-N consistency. In general this will require restrictions on the rate of increase of K.

[4]See Rothenberg (1983) for example. Also note that the expression for the bias is the same as Nagar (1959) when $a = 0$.

The result in Proposition 1 does not reveal the large K approximate MSE for either LIML or B2SLS because neither of these estimators have a bias (to $O(N^{-1/2})$) that depends on K. Thus we need to provide an additional result that will yield approximate MSE's for such estimators.

Proposition 2 *Given Assumptions 1-3, assuming that* $a = O(1)$ *and* $b = O(K)$ *such that* $(1-a)K - b = O(1)$, *the MSE for* $\sqrt{N}(\hat{\gamma} - \gamma)$ *is approximately*

$$\sigma_{u\epsilon}^2 \frac{(K - 2aK - 2ab)}{N} + \sigma_u^2 \sigma_\epsilon^2 \frac{K}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}.$$

This result can be used to obtain approximate MSE's for LIML and B2SLS.

Corollary 2 *For LIML,* $\hat{\gamma}_L$, *which has* $a = 1$ *and* $b = 0$, *the approximate MSE for* $\sqrt{N}(\hat{\gamma}_L - \gamma)$ *under Assumptions 1-3, is given by,*

$$(\sigma_u^2 \sigma_\epsilon^2 - \sigma_{u\epsilon}^2)\frac{K}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}$$

Corollary 3 *For B2SLS,* $\hat{\gamma}_B$, *which has* $a = 0$ *and* $b = K - 2$, *the approximate MSE for* $\sqrt{N}(\hat{\gamma}_B - \gamma)$ *under Assumption 1-3, is given by,*

$$(\sigma_u^2 \sigma_\epsilon^2 + \sigma_{u\epsilon}^2)\frac{K}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}.$$

The expressions derived thus far do not allow one to derive a criteria for JIVE. Interestingly, the following result shows that JIVE has the same MSE expression as B2SLS.

Proposition 3 *For JIVE,* $\hat{\gamma}_J$, *the approximate MSE for* $\sqrt{N}(\hat{\gamma}_J - \gamma)$ *is given by,*

$$(\sigma_u^2 \sigma_\epsilon^2 + \sigma_{u\epsilon}^2)\frac{K}{N} + \sigma_\epsilon^2 \frac{f'(I-P)f}{N}$$

3.2 Discussion

Some insight into the differences between the approximate MSE for the estimators can be gained by looking at the terms in the expansion of each that contributes to the leading terms in the MSE expressions. For 2SLS this term has the form,

$$\frac{u'P\epsilon}{\sqrt{N}} = \frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N} u_i\epsilon_i P_{ii} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i\epsilon_j P_{ij}\right) \tag{2}$$

where the first term on the right reflects part of the bias that occurs due to the endogeneity in the model (i.e. correlation of u_i and ϵ_i) and which grows with K.[5] The square of this bias term gives the leading term in the MSE for 2SLS. The variance of the complete expression contributes $O(K/N)$ terms to the MSE, and these are dominated by the leading bias squared term. Both LIML and B2SLS have the leading bias term (which depends on K) removed and so the leading terms in their MSE come from the variance terms. An alternative derivation of LIML shows how this bias reduction occurs. In particular the LIML estimator $\hat{\gamma}_L$ can be shown to be the solution to the following problem,

$$\min_{\gamma} \frac{\gamma_*'\bar{Y}P\bar{Y}\gamma_*}{\gamma_*'\bar{Y}\bar{Y}\gamma_*}$$

where $\bar{Y} = (y, Y)$, and $\gamma_*' = (1, -\gamma)$. Given this, it is straightforward to show that the LIML estimator satisfies the following first order condition,

$$\frac{1}{\sqrt{N}}\left(Y'P\epsilon(\hat{\gamma}_L) - \frac{Y'\epsilon(\hat{\gamma}_L)}{\epsilon(\hat{\gamma}_L)'\epsilon(\hat{\gamma}_L)}\epsilon(\hat{\gamma}_L)'P\epsilon(\hat{\gamma}_L)\right) = 0$$

where $\epsilon(\hat{\gamma}_L) = y - \hat{\gamma}_L Y$. Letting,

$$\hat{\delta}(\hat{\gamma}_L) = \frac{Y'\epsilon(\hat{\gamma}_L)}{\epsilon(\hat{\gamma}_L)'\epsilon(\hat{\gamma}_L)}$$

we can rewrite the first order condition as,

$$\frac{1}{\sqrt{N}}(Y - \hat{\delta}(\hat{\gamma}_L)\epsilon(\hat{\gamma}_L))'P\epsilon(\hat{\gamma}_L) = 0. \tag{3}$$

[5]Note that $\sum_{i=1}^{N} P_{ii} = K$ since P is a projection matrix with rank K.

Since $\hat{\delta}$ can be interpreted as being the coefficient in the linear regression of Y on ϵ the first term in the above can be viewed as being the residual from this regression, which should then not be linearly related to ϵ. Thus the term for LIML that corresponds to (2) has the form,

$$\frac{1}{\sqrt{N}} v' P \epsilon = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{N} v_i \epsilon_i P_{ii} + \sum_{i=1}^{N} \sum_{j \neq i}^{N} v_i \epsilon_j P_{ij} \right) \tag{4}$$

where,

$$v_i = u_i - \frac{\sigma_{u\epsilon}}{\sigma_\epsilon^2} \epsilon_i$$

is by construction uncorrelated with ϵ_i. Note also that the leading term in the MSE expression for LIML is equivalent to, $\sigma_v^2 \sigma_\epsilon^2$. The term for B2SLS corresponding to (2) and (4) is equal to,

$$\frac{u' P \epsilon - K \sigma_{u\epsilon}}{\sqrt{N}} = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{N} (u_i \epsilon_i P_{ii} - K \sigma_{u\epsilon}) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} u_i \epsilon_j P_{ij} \right) \tag{5}$$

where the leading term has, by construction, zero expectation. This, again, results in a lower order leading term in the MSE for B2SLS. The lower leading term for LIML compared to that of B2SLS comes about because of the lower variance that arises because v_i is uncorrelated with ϵ_i and because v_i has a lower variance than u_i. Finally, it is easy to see why JIVE has the same leading term as B2SLS, since for JIVE the expression that is analogous to (2), (4) and (5) has the form,

$$\frac{u' C \epsilon}{\sqrt{N}} = \frac{1}{\sqrt{N}} \left(\sum_{i=1}^{N} \sum_{j \neq i}^{N} u_i \epsilon_j C_{ij} \right) \tag{6}$$

which has zero expectation. This expression results because of the use of the Jackknife in the first stage.

3.3 Further Comparisons

Our discussion thus far has indicated that, in terms of dominant terms in the approximate (large K) MSE, LIML dominates JIVE and B2SLS and that all of these estimators dominate 2SLS because of the larger order (in K) leading term of the latter. Another

type of comparison that can be made is in terms of the MSE for a fixed value of K for which, generally speaking, the chosen instrument set may be suboptimal. The nice feature of this comparison, is that if the expansion is taken up to terms of order N^{-1} then one can see clearly the opposing effects of increasing K more clearly. In addition such an expansion extends the work of Nagar (1959) and Rothenberg (1983) to the misspecified first stage case and may reveal any potential differences between estimators when one uses an imperfect fixed set of fixed instruments in the first stage.[6]

Proposition 4: *Given Assumptions 1, 2(i), 3, assuming that a, b and K are fixed and that (for some small constant c),*

$$0 < c < \frac{f'Pf}{N} < c^{-1}$$

with probability 1, uniformly in N, then the MSE for $\sqrt{N}(\hat{\gamma} - \gamma)$ using terms up to those of $O(N^{-1})$ is approximately

$$\begin{aligned}
&\sigma_\epsilon^2 \left(\frac{f'Pf}{N}\right)^{-1} + \sigma_{u\epsilon}^2 \frac{[(1-a)K-b]^2}{N} \left(\frac{f'Pf}{N}\right)^{-2} \\
&+\sigma_{u\epsilon}^2 \left(\frac{\{(K-2aK-2ab) - 8[(1-a)K-b]\}}{N}\right) \left(\frac{f'Pf}{N}\right)^{-2} \\
&+\sigma_u^2\sigma_\epsilon^2 \left(\frac{\{K - 2[(1-a)K-b]\}}{N}\right) \left(\frac{f'Pf}{N}\right)^{-2} \\
&+\sigma_{u\epsilon}^2 \left(\frac{12-4a-a^2}{N}\right) \left(\frac{f'Pf}{N}\right)^{-2} + \sigma_u^2\sigma_\epsilon^2 \left(\frac{4-2a}{N}\right) \left(\frac{f'Pf}{N}\right)^{-2} \\
&+2\sigma_\epsilon^2 \left(\frac{a(K-1)+b}{N}\right) \frac{f'(I-P)f}{N} \cdot \left(\frac{f'Pf}{N}\right)^{-2}
\end{aligned}$$

The conditions used to derive the expression in Proposition 4 are the same as used previously except that we have introduced a condition that modifies Assumption 2(ii) to

[6]Indeed our result reduces to that of Nagar (1959) in the perfect first stage case when $a = 0$ as he assumed. The expression obtained in Proposition 4 does differ from the MSE expression one obtains from Theorem 2 of Rothenberg (1983). Additionally, in the $a = 0$ case, the expression one obtains from Rothenberg (1983) is not consistent with that obtained by Nagar (1959). The difference between that obtained from Rothenberg (1983) and Nagar (1959) and ours is minor and appears appear to be due to typographical error in the statement of Theorem 2 of Rothenberg (1983). Indeed a single sign change results in all espressions being consistent with one another.

ensure identification in the fixed K case. Note that the leading term in this expression, when taken to the limit, is the asymptotic variance of the estimator, when K is fixed. When K is fixed and when the instruments are imperfect (in the sense that they cannot predict f perfectly) then it is the case that,

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N} \right)^{-1} \geq \sigma_\epsilon^2 \left(\frac{f'f}{N} \right)^{-1}$$

since $f'(I-P)f \geq 0$ which follows because $I-P$ is positive semidefinite. In fact if the instruments are imperfect (in the sense that they cannot predict f perfectly) then it is the case that the inequality is strict. Note that the expression on the right of the inequality is the lowest variance possible for this particular problem. Given the expression in Proposition 4, one can then see clearly that increasing K will have opposing effects. First, increasing K will cause the limiting variance to be closer to the optimal approximate variance for the estimator. That is,

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N} \right)^{-1} - \sigma_\epsilon^2 \left(\frac{f'f}{N} \right)^{-1} \to 0$$

because when $K \to \infty$ then by Assumption 2, $f'(I-P)f/N \to 0$. On the other hand the other terms will generally increase. Our expressions derived earlier, involve the largest terms reflecting each of these effects. The optimal choice of K , then, will be the one that balances these two opposing effects.

The expression in Proposition 4 can also be used to compare the different estimators in the class under consideration. In particular the approximate MSE for 2SLS is,

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N} \right)^{-1} + \left\{ \sigma_{u\epsilon}^2 \left(\frac{K^2 - 7K + 12}{N} \right) + \sigma_u^2 \sigma_\epsilon^2 \frac{4-K}{N} \right\} \left(\frac{f'Pf}{N} \right)^{-2}$$

for LIML we have,

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N} \right)^{-1} + \left((\sigma_u^2 \sigma_\epsilon^2 - \sigma_{u\epsilon}^2) \frac{K}{N} + \left(2\sigma_{u\epsilon}^2 - \sigma_u^2 \sigma_\epsilon^2 \right) \frac{1}{N} \right) \left(\frac{f'Pf}{N} \right)^{-2}$$
$$+ 2\sigma_\epsilon^2 \left(\frac{(K-1)}{N} \right) \frac{f'(I-P)f}{N} \cdot \left(\frac{f'Pf}{N} \right)^{-2}$$

and finally for B2SLS we have,

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N}\right)^{-1} + \left\{(\sigma_u^2\sigma_\epsilon^2 + \sigma_{u\epsilon}^2)\frac{K}{N} + 2\sigma_\epsilon^2 \left(\frac{K-2}{N}\right)\frac{f'(I-P)f}{N}\right\} \cdot \left(\frac{f'Pf}{N}\right)^{-2}$$

To allow comparisons of these k-class estimators with JIVE in the fixed K case we derive the appropriate MSE expression for JIVE below in Proposition 5.

Proposition 5: *Given Assumptions 1, 2(i), 3, assuming that a, b and K are fixed,* $\sup_i P_{ii} = O(N^{-1})$ *and that,*

$$0 < c < \frac{f'C'f}{N} < c^{-1}$$

with probability 1, uniformly in N, then the MSE for $\sqrt{N}(\hat{\gamma}_J - \gamma)$ using terms up to those of $O(N^{-1})$

$$\sigma_\epsilon^2 \left(\frac{f'Pf}{N}\right)^{-1} + (\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2)\frac{K}{N}\left(\frac{f'P'f}{N}\right)^{-2} + \left(12\sigma_{u\epsilon}^2 + 4\sigma_u^2\sigma_\epsilon^2\right)\frac{1}{N}\left(\frac{f'P'f}{N}\right)^{-2} + 2\sigma_\epsilon^2 \left(\frac{f'(I-P)\tilde{P}(I-P)f}{N}\right)\left(\frac{f'P'f}{N}\right)^{-2}$$

Note the additional condition in the statement of the result which concerns the rate at which the largest diagonal of the matrix P goes to zero which is a strengthening of Assumption 3 and is not needed for any other result. It is fairly easy to provide conditions that justify this condition in the fixed K case considered in the result – for instance given that Assumption 2 holds one can verify this condition if the smallest eigenvalue of the matrix $\Psi'\Psi/N$ is bounded away from zero. The reason for imposing this condition is to get an estimate of,

$$\left(\frac{f'Pf}{N}\right)^{-1} - \left(\frac{f'C'f}{N}\right)^{-1}\frac{f'C'Cf}{N}\left(\frac{f'C'f}{N}\right)^{-1}$$

where the complicated form for JIVE comes from the use of the Jackknife in the first stage, and similar to the results for LIML and B2SLS, introduces an additional $O(N^{-1})$, term in the MSE. Such an additional term does not appear in the expression for 2SLS. Thus any comparison in the fixed K case will depend not only on K, and the values for

the variances and covariances but also on the degree of misspecification of the first stage. It is interesting to note that although this misspecification obviously affects all estimators through the leading term (the approximate variance) it also affects LIML, B2SLS and JIVE in the smaller order terms. Indeed one could envision scenarios where K is fixed at some large number and where there is still a great deal of misspecification in the first stage - in such a scenario it may be possible for the final term in the MSE for LIML, B2SLS and JIVE to be large and indicate poor behavior of these estimators relative to 2SLS.

One can do more general comparisons of the estimators in the case where there is a correct specification in the first stage so that $Pf = f$. In this case the last terms in the MSE for LIML, B2SLS and JIVE would be omitted. Comparing the resulting MSE reveals that although JIVE and B2SLS have similar bias reduction properties and the same criteria for picking K, B2SLS appears to be superior with respect to terms that do not depend on K (at least using the approximate MSE to $O(N^{-1})$). A comparison of the MSE for LIML and B2SLS reveals that LIML has a lower MSE. The comparison of 2SLS with the other estimators depends on the variances and covariances in the model. Not surprisingly large (fixed) values of K and reasonable levels of covariance between the residuals would tend to lead to superior MSE for the other estimators relative to 2SLS.

The expressions in Propositions 4 and 5 also makes it possible to consider what happens when there is no endogeneity problem, so that $\sigma_{u\epsilon} = 0$. When $\sigma_{u\epsilon} = 0$, the optimal number of instruments for JIVE and B2SLS, using the criteria developed in Section 3.2, is exactly the same as for LIML, which is equal to the optimal number of terms required for estimating f in the first stage. On the other hand, the MSE expression just given for 2SLS, (when $\sigma_{u\epsilon} = 0$) is monotonically decreasing in K which indicates that the optimal choice of K is as large as possible. This, not surprisingly, makes 2SLS identical to OLS, and is due to the fact that when $\sigma_{u\epsilon} = 0$, OLS is the Best Linear Unbiased Estimator. The reason that this does not occur with LIML is that the LIML estimator would no longer be defined when there are as many instruments as observations while the delete-one nature of the first stage used in JIVE makes first stage predictions

based on as many instruments as observations very unreliable.

Because of the dependence of comparisons on the degree of correlation between residuals, it is difficult to use the expressions in Propositions 2 and 4 to derive an convenient optimal estimator in the class under consideration. It is possible, however to describe the optimal estimator for classes of estimators that have similar bias properties.[7] The first result will use the expression in Proposition 2 to give a class of estimators that are optimal with respect to the large sample large K MSE to order K/N. The class under consideration consists of those estimators whose bias is of $O(N^{-1/2})$ (and hence have a squared bias of $O(N^{-1})$) which, as noted earlier, places restrictions on a and b that rule out 2SLS but include LIML and B2SLS as possibilities. This result is stated below in Lemma 1 and is an immediate consequence of Proposition 2.

Lemma 1 *In the case where $K \to \infty$ in such a way that $(1-a)K-b = O(1)$, the optimal estimator, with respect to the large sample (large K) MSE to $O(K/N)$ must have $a = 1$ and $b = O(1)$.*

Note that this result does not pin down a single estimator, since one can pick any constant value for b and still have the same leading terms in the MSE expansion, since terms involving K/N will dominate those of $O(N^{-1})$ for large K. It does suggest, however, that LIML, being one such estimator, will have better (large K) MSE properties than the other estimators under consideration. Note, again, that 2SLS does not fit into the class of estimators under consideration because its bias is not $O(N^{-1/2})$ (again as $K \to \infty$) as the condition in the result requires. The next result is for the class of estimators that are unbiased to $O(N^{-1/2})$ and follows similarly from Proposition 2.

Lemma 2 *Restricting attention to estimators that are such that $(1-a)K-b-(2-a) = 0$, the optimal estimator, with respect to the large sample large K MSE to $O(K/N)$ must have $a = 1$ and consequently $b = 1$.*

[7]Indeed this was done by Rothenberg (1983) for the class of (approximately) median unbiased estimators, in which LIML was optimal.

This result states that the optimal estimator in the class of estimators that are unbiased to $O(N^{-1/2})$ is the bias adjusted LIML estimator. It should be noted, however, that for such an estimator the optimal choice of K is completely unaffected by the bias adjustment and that the criteria for choosing K will be identical to that obtained for LIML. A similar result to that in Lemma 2 is obtained when one assumes that K is fixed and considers approximately mean unbiased estimators with MSE given by the expression in Proposition 4.

Lemma 3 *Restricting attention to estimators that are such that* $(1-a)K-b-(2-a) = 0$, *the optimal estimator, with respect to the large sample fixed* K *MSE to* $O(N^{-1})$ *must have* $a = 1$ *and consequently* $b = 1$.

This result shows that when there is misspecification in the first stage (as happens in general with K fixed) the bias adjusted LIML estimator is optimal with respect to the MSE to $O(N^{-1})$. This is a minor extension, to the misspecified first stage case, of the conclusions one draws from Rothenberg (1983) concerning the optimality of the bias adjusted LIML estimator among the class of approximately mean unbiased estimators. In addition to this, Rothenberg (1983) showed that for fixed K and a correctly specified first stage LIML is optimal (with respect to the MSE to $O(N^{-1})$) among all (approximately) *median* unbiased estimators. Thus based on both the dominant terms in large K MSE and peripheral or second order terms there appears to be ample evidence to suggest the use of LIML or bias adjusted LIML.

3.4 Admitting Covariates

In this section we generalize the results to the model,

$$y_i = \gamma Y_i + x'_{1i}\beta + \epsilon$$

where x_{1i} is a $d_1 \times 1$ subvector of x_i. The most important difference that arises in the more general formulation is that we are now concerned with a vector of parameters, $(\gamma, \beta'_1)'$.

However, it can be shown that the MSE of estimators of β_1 are directly and positively related to the MSE of estimators of γ. In particular, assuming that x_{1i} is included in the instrument set (which would be the case if one were using power series) then any estimator of $\beta_1$will have the form,

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - \hat{\gamma}Y)$$

where X_1 denotes the $N \times d_1$ matrix with ith row qual to x_{1i}. [8] Since for all three estimators the MSE of $\hat{\beta}_1$ depends on K only through its dependence on $\hat{\gamma}$, we can focus directly on the MSE of the $\hat{\gamma}$ in each case. When covariates are present the estimators of γ take the form,

$$\hat{\gamma} = ((1+\kappa)Y^{*\prime}PY^* - \kappa Y^{*\prime}Y^*)^{-1}((1+\kappa)Y^{*\prime}Py - \kappa Y^{*\prime}y)$$

where $Y^* = (I - P_1)Y$ with P_1 being the projection matrix formed using X_1. The JIVE estimator will take the form,

$$\hat{\gamma}_J = (Y'C'(I - P_1)Y)^{-1}Y'C'(I - P_1)y.$$

Before presenting the MSE calculations we modify Assumption 2 as follows, letting π denote the population regression coefficients from the regression of Y_i on x_{1i}.

Assumption 2' *The function $f(x_i) = E(Y_i|x_i)$ is such that,*

(i) $f : X \to R$ and has an extension to all of R^d that is $s > 0$ times continuously differentiable in x and,

(ii) with probability 1 $0 < c < N^{-1}\sum_{i=1}^{N}(f(x_i) - x_{1i}'\pi)^2 < c^{-1}$ for some small constant c uniformly in N

(iii) the smallest eigenvalue of $X_1'X_1/N$ is bounded away from zero uniformly in N with probability one.

[8] Note that this is true of JIVE as well since $CX_1 = X_1$.

The adjustments in Assumption 3' are due to the fact that the identification condition now suggests that x_i explain y_i beyond that provided by a linear regression on x_{1i}. [9] The new condition (iii) along with Assumption 1(v) is used to bound the elements of the vector, $(I-P_1)f$ and the variances and covariances of elements of $(I-P_1)u$ and $(I-P_1)\epsilon$. Indeed these conditions ensure that the variance of a typical element of $(I-P_1)u$, say $u_i - P_{1i}'u$ (with P_{1i} denoting column i of P_1), is equal to that of u_i plus a term that is $O(N^{-1})$ and that any covariances between these elements are of second order importance. The same holds true for the vector $(I-P_1)\epsilon$ and for covariances between these vectors. Once this has been noted one can obtain large K MSE expressions for the estimators of interest that are the same as in the case of no covariates provided that the set basis functions used as instruments contain linear terms in x_{1i}.

Corollary 4 *Given Assumptions 1,2' and 3 plus $PP_1 = P_1$ the conclusions of Corollaries 1, 2 and 3 and Proposition 3 all hold.*

It is easy to see why this result holds when $PP_1 = P_1$. The only difference that would obtain by allowing for exogenous covariates in the equation would be that the second terms in the approximate MSE would have the form,

$$f'(I-P_1)(I-P)(I-P_1)f/N = f'(I-P)f$$

where the equality comes from the fact that $PP_1 = P_1$. A simple consequence of this result is that rules for choosing K and any optimality results discussed previously apply carry over to the situation with covariates.

4 Feasible Optimal Estimation of K

In this section we consider the properties of the estimated MSE criteria discussed in Section 2. Throughout this section we will use the following assumptions concerning the preliminary estimation of the parameters, σ_u^2, σ_ϵ^2 and $\sigma_{u\epsilon}$.

[9] This is essentially a relevance condition and Donald (1997) has considered tests for the failure of this condition in contexts where there are an arbitrary number of right hand side endogenous variables.

Assumption 4 *Assume the following,*

(i) $\hat{\sigma}_u^2 \xrightarrow{p} \sigma_u^2$,

(ii) $\hat{\sigma}_\epsilon^2 \xrightarrow{p} \sigma_\epsilon^2$

(iii) $\hat{\sigma}_{u\epsilon} \xrightarrow{p} \sigma_{u\epsilon}$.

As noted in Section 2, this can be achieved by using some preliminary set of estimates of the model obtained by using the number of instruments that provide the optimal fit in the first stage, based on either the cross-validation or Mallows criteria. Note that the other key ingredient used to form the estimated criteria is $\hat{R}(K)$ which can either be based on cross validation or the Mallows criteria as discussed in Section 2. Li (1987) has shown that each of these criteria are asymptotically optimal in the sense that the value of K that minimizes these criteria, denoted $\hat{K}$, satisfies the condition,

$$\frac{R(\hat{K})}{\inf_K R(K)} \longrightarrow 1$$

in probability.[10] For a generic estimator $\hat{\gamma}$ with estimated criteria $\hat{S}(K)$ and true criteria $S(K)$ we will define the data driven optimal choice of the number of instruments by,

$$\hat{K} = arg \min_K \hat{S}(K).$$

The minimum and all infimum will be relative to an index set of K values, which may require some restrictions in order to prove certain results. Also, notice that although $\hat{R}(K)$ only estimates $R(K)$ up to a constant, which does not depend on K, one can remove this constant from $\hat{R}(K)$ and hence from $\hat{S}(K)$ without affecting the choice of K. To analyze the properties of $\hat{K}$ we follow the approach of Li (1987) and Andrews (1991b). In our case we use the following definition of optimality.

Definition: *A method of selecting $\hat{K}$ is defined to be "higher order asymptotically optimal with respect to the criteria $S(K)$" if it can be shown that,*

$$\frac{S(\hat{K})}{\inf_K S(K)} \xrightarrow{p} 1.$$

[10] Andrews (1991) proved the same results for appropriate adaptations of these criteria in the case where the residuals are heteroskedastic.

The use of the term "higher order" is necessitated by the fact that the optimal MSE convergence rate in all cases will be N^{-1} (as $N \to \infty$) which cannot be improved upon, and the criteria based on $S(K)$ concerns the next largest terms in the MSE that depend on the choice of K. A vital step in showing that the criteria are higher order asymptotically optimal, is showing that $\hat{R}(K)$ consistently estimates $R(K)$ up to the constant in the sense that, in probability,

$$\sup_K \frac{|\hat{R}(K) - \bar{\sigma}_u^2 - R(K)|}{R(K)} \to 0. \tag{7}$$

This is the condition used by Li (1987) to show the asymptotic optimality of the Mallows criteria and cross validation. In order to be able to verify this result we employ conditions similar to those of Li (1987).

Assumption 5: *Assume that the following conditions hold,*

(i) $E(u_i^8|x_i) < \infty$ *and*

(ii) $\inf_K NR(K) \to \infty$,

and in addition when $\hat{R}(K)$ *is the cross validation criteria, assume*

(iii) the index set satisfies the condition that $\sup_K \sup_i P_{ii} \to 0$ *in probability.*

The conditions (i) and (ii) in Assumption 5 are needed for both the Mallows criteria and cross validation, to show that they satisfy (7). These were employed by Li (1987) to show that these criteria are asymptotically optimal. Here, this is an intermediate step to showing higher order asymptotic optimality. The second of these requires either that f not have a finite order representation in terms of the basis functions or else that the lower bound on the set of K values over which one is doing the minimization grows with N, which is perhaps undesirable from a practical standpoint. The condition (iii) is required only for the cross validation based method, and is similar to the condition in Assumption 3, which was used to derive the MSE criteria. This condition will essentially place an

upper bound on the rate at which the index set can grow with the sample size [11]. The following results show that for each estimator the estimated criteria provides a means to obtain higher order asymptotically optimal choices for the number of instruments.

Proposition 6: *Given Assumptions 4 and 5, then the rule "select the value K that minimizes $\hat{S}^L(K)$" is higher order asymptotically optimal with respect to the criteria $S_L(K)$.*

Proposition 7: *Given Assumptions 4 and 5, then the rule "select the value K that minimizes $\hat{S}_2(K)$" is higher order asymptotically optimal with respect to the criteria $S_2(K)$.*

Proposition 8: *Given Assumptions 4 and 5, the rule "select the value K that minimizes $\hat{S}_J(K)$" is higher order asymptotically optimal with respect to the criteria $S_J(K)$.*

5 Simulation Study

In this section we report the results of a small Monte-Carlo experiment which has been designed along the lines of that used in Angrist, Imbens and Krueger (1995) (hereafter AIK). Three basic designs are used. All experiments are based on the estimation of the equation,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \tag{8}$$

where $(\beta_0, \beta_1) = (0, 1)$ and where X_{i1} is an endogenous explanatory variable. Three cases are considered and are distinguished by differences between X_{i1} and a set of potential instruments.

Case 1: In this case X_{i1} is related to either $k = 10$ or $k = 20$ independent standard normal random variables, Z_{ij} through the linear equation,

$$X_{i1} = \pi_0 + \sum_{j=1}^{k} \pi_j Z_{ij} + \eta_i \tag{9}$$

[11] As noted below Assumption 3, based on Donald (1997) has shown, using the results in Newey (1995), that for power series, if x_i has a continuous distribution with density that is bounded and bounded away from 0 on X then the upper bound on the index set of K values, say $\bar{K}$, should satisfy, $\bar{K}^5/N \to 0$.

with $\pi_1 = 0.3$ and $\pi_j = 0$ for $j \neq 1$, so that only the first Z_{ij} is relevant for predicting X_{i1}. In this case the residuals in the two equations are generated as,

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}\right)$$

This case is identical to Model 2 of AIK, who used a complete set of $k = 20$ instruments in their experiments.

Case 2: This is the same as in Case 1, except that now in (9) $\pi_5 = 0.3$ and $\pi_j = 0$ for $j \neq 5$ so that only the fifth instrument is relevant. Again experiments are conducted for both $k = 10$ and $k = 20$ potential instruments (in addition to the constant). The errors are generated in the same way as in Case 1.

Case 3: This is the same as Model 3 of AIK and specifies,

$$X_{i1} = \pi_0 + \sum_{j=1}^{k} \pi_j Z_{ij} + 0.3 \sum_{j=2}^{k} Z_{ij}^2 + \eta_{i0} \sum_{j=2}^{k} Z_{ij}^2/19 \tag{10}$$

with $\pi_1 = 0.3$ and $\pi_j = 0$ for $j \neq 1$, so that in addition to the first regressor being relevant the others enter in a nonlinear fashion. Additionally, the errors in the first stage are heteroskedastic. The selection process will, however, only consider using the Z_{ij} variables themselves, and will consider either up to the first $k = 10$ instruments or the first $k = 20$ instruments. The residuals in the pair of equations (8) and (10) are generated as,

$$\begin{pmatrix} \epsilon_i \\ \eta_{i0} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.00 & 0.80 \\ 0.80 & 1.00 \end{pmatrix}\right)$$

For each of the three cases, experiments are conducted with samples of size $N = 100$ and $N = 400$, and using a maximum number of instruments of $k = 10$ and $k = 20$. Thus for each case a total of four different experiments were conducted. The number of replications was set at $5,000$. In each experiment the 2SLS, LIML and JIVE estimators were obtained using the number of instruments that minimized the objective functions discussed in the previous two sections. These criteria were constructed using delete-one cross validation of the first stage relationship, along with estimates of $\sigma_{u\epsilon}^2$, σ_u^2 and σ_ϵ^2 which were obtained as discussed in Section 4, using a preliminary estimator which used all of the instruments.

Tables 1.1, 2.1 and 3.1 contain the relevant summary statistics for each of the three cases respectively. In each table we report the median of the 5,000 estimates along with the 0.1 and 0.9 quantiles of the estimates. The difference between these can be considered a way of measuring dispersion. Additionally we report the Mean Absolute Error of the estimates, which is an alternative measure of dispersion, and, finally, the coverage rate of a nominal 95 percent confidence interval for each estimator. Additionally we report the relevant statistics for the OLS estimator as well as 2SLS and LIML estimators that use all $k = 20$ instruments. These statistics were reported in AIK (1995, Tables 2 and 3). In Tables 1.2, 2.2 and 3.2 we report the frequencies with which the various possible K values were chosen in the experiments.

A few features of the results are worth noting. First, the most dramatic improvements occur in the use of 2SLS with the optimal number of instruments being used. As indicated in Tables 1.1 and 3.1, 2SLS which uses all the instruments has very poor properties, and is biased towards OLS. Moreover, confidence intervals generally have a very poor coverage rate with the worst occurring in Case 1 with $N = 100$ where the nominal 95% confidence interval contains the true value just 22% of the time. The other cases are similarly poor. When the criteria are used to choose the number of instruments there is a substantial reduction in the bias (as measured by the median) and the coverage rates for the confidence intervals are much closer to the nominal rate of 95% generally being around 90%. The only exception to this is Case 2 with $N = 100$ where the coverage rate is below 80%, although this is a dramatic improvement from the 23% coverage rate when all instruments are used. The other noticeable difference in 2SLS when the criteria is used is that the dispersion is much lower as evidenced by reduced MAE and a lower spread between the 10% and 90% quantiles. It also worth noting that the performance of 2SLS using the criteria is not sensitive to the maximum K being considered. Indeed the results for 2SLS when one uses the criteria to pick K with a maximum possible K of 10 are almost identical to those when the maximum possible K is 20.

For LIML on the other hand, the main improvements relative to the estimator that uses all the instruments appears to be in terms of reduced variability, either as measured

by MAE or as the difference between the .1 and .9 quantiles. This is because LIML performs fairly well when all of the instruments are used and is approximately median unbiased and provides confidence intervals that have a coverage rate that is reasonably close to the nominal rate of 95% being around 90%. The coverage rate for LIML when the criteria is used is slightly lower than for LIMLALL in Cases 1 and 2. In Case 3, however, there is an improvement in the coverage rate when the criteria is used with the rate being very close to the nominal rate. The main improvement that comes from the use of the criteria is in the decreased dispersion of the estimator which results in all cases.

The performance of JIVE is a little bit mixed relative to the other estimators. The bias (using the median) of JIVE in some cases appears to be the lowest among the estimators and other cases (notably Case 3 with 100 observations) is the worst so that there is no clearly superior estimator as far as bias is concerned. What is striking is that the dispersion for JIVE, using either measure, is larger in practically all cases relative to either the LIML or 2SLS estimators that result from the use of the criteria for picking the number of instruments. There is also no clear improvement in the coverage rates from the use of JIVE rather than either 2SLS or LIML estimators that use the criteria for picking the number of instruments.

In Tables 1.2, 2.2 and 3.2 we have provided frequencies with which the criteria chose the respective number of instruments for each estimator in each case. A few features stand out. First, for all estimators the criteria usually points to a value of K that is at least as large as is required to include the relevant instrument – in Cases 1 and 3 this is $K = 1$ while in Case 2 this is $K = 5$. What is also interesting, is that in Case 2 the criteria generally pick the correct number of instruments more than half the time. Also, as one might expect from the discussion in Section 3, the criteria for 2SLS generally points to smaller values of K than does the criteria for JIVE, while the latter generally picks values for K that are smaller (on average) than those for LIML. With LIML and JIVE the criteria do with some positive probability point to an instrument set that includes many irrelevant instruments. Fortunately, as evidenced by the other sets of results, this does not cause much of a problem for either estimator.

6 Application to the Returns to Schooling

Angrist and Krueger (1991) in an important paper used the quarter of birth as an instrument for the number of years of education – since laws typically dictate that people should stay in school until they reach a certain age, one would think that for students that drop out once they reach the age, the amount of education they receive will depend to some extent on when in the year they were born. Using a sample drawn from the 1980 U.S. Census that consisted of 329,500 men born between 1930-1939, Angrist and Krueger (1991) estimate an equation where the dependent variable is the log of the weekly wage, and the explanatory variable of interest is the number of years of schooling. The set of excluded variables that were used as instruments included 30 variables that were the result of interacting dummies for the year of birth with dummies for the quarter of birth, plus another 150 variables that were the result of interacting the quarter of birth dummy with dummies for the state in which the individual was born. Thus a total of 180 instruments were used. See Angrist and Krueger (1991) for more details.

Using OLS the returns to education (the coefficient of the schooling variable) was 0.067 with a standard error of 0.0003. Using 2SLS Angrist and Krueger (1991) obtained 0.093 and a standard error of 0.009. Table 4 contains various estimates for different instrument sets. These instrument sets consist of the set of 30 instruments constructed from interacting the quarter of birth and year of birth dummies plus the various sets indicated in the table. These are a number of subsets of regional dummies, representing various partitions of the U.S. up to and including the set of 50 state of birth dummies. Also included as potential instruments is the interaction of these with the quarter of birth dummies. Also indicated in the Table are the instrument selection criteria for 2SLS and LIML. As can be seen the criteria indicate that only a very small set of the dummies should be used – for 2SLS, only the 4 region dummies should be used, while LIML seems to require that these plus their interactions with the quarter of birth dummy be used as instruments. It is interesting to note that a straight cross validation measure for the relationship between schooling and these dummies also indicates that the set of dummies

that is optimal for LIML is also optimal for this first stage relationship (at least among the sets of estimates used).

Although the estimates obtained from both 2SLS and LIML are not very different from what one obtains when one uses all the instruments, it is noteworthy that the standard error for 2SLS is somewhat larger with the smaller set of instruments. Moreover, it seems to be the case that the standard error is close to that of LIML, and that LIML's standard error is less sensitive to the inclusion of useless instruments.

7 Appendix A

Proof of Proposition 1: For large enough K the expression for the MSE presented in Proposition 4 remains valid by Assumption 2 which ensures that for large enough K the term $f'Pf/N$ is invertible. Take this expansion which, as shown in the proof of Proposition 4, has a remainder term that is $o(K^2/N)$, and note that since we are finding the MSE of $H\sqrt{N}(\hat{\gamma}-\gamma)$ we need to multiply each term in the expansion by H^2 (noting that under Assumption 2(ii) $H = O(1)$). Since $K \to \infty$ and that since $a = O(1)$, $b = O(K)$ and $(1-a) - b/K > 0$ then the leading term among those that increase with K will be the term,

$$\sigma_{u\epsilon}^2 \frac{[(1-a)K-b]^2}{N} \left(\frac{f'Pf}{N}\right)^{-2} H^2 = O(K^2/N) \tag{11}$$

Also note that if $K \to \infty$ then the term

$$\sigma_{\epsilon}^2 \left(\frac{f'Pf}{N}\right)^{-1} \tag{12}$$

decreases with K. Use the expansion,

$$\frac{f'Pf}{N} = \frac{f'f}{N} - \frac{f'(I-P)f}{N}$$

and the fact that under $K \to \infty$, $f'(I-P)f/N = o(1)$ by Assumption 2, so that,

$$\begin{aligned} \left(\frac{f'Pf}{N}\right)^{-1} &= \left(\frac{f'f}{N}\right)^{-1} + \frac{f'(I-P)f}{N}\left(\frac{f'f}{N}\right)^{-2} + O\left(\left[\frac{f'(I-P)f}{N}\right]^2\right) \\ &= H^{-1} + \frac{f'(I-P)f}{N}H^{-2} + O\left(\left[\frac{f'(I-P)f}{N}\right]^2\right) \end{aligned} \tag{13}$$

Now, substitute this expansion into the (11) and (12) to obtain respectively,

$$\sigma_{u\epsilon}^2 \frac{[(1-a)K-b]^2}{N} \left(\frac{f'Pf}{N}\right)^{-2} H^2 = \sigma_{u\epsilon}^2 \frac{[(1-a)K-b]^2}{N} + o(\frac{K^2}{N})$$

and,

$$\sigma_{\epsilon}^2 \left(\frac{f'Pf}{N}\right)^{-1} H^2 = \sigma_{\epsilon}^2 H + \sigma_{\epsilon}^2 \frac{f'(I-P)f}{N} + O\left(\left[\frac{f'(I-P)f}{N}\right]^2\right) \tag{14}$$

and take the dominant terms among those that increase with K and those that decrease with K. **Q.E.D.**

Proof of Proposition 2: The proof is similar to that of Proposition 1, except that after multiplying the expression in Proposition 4 by H^2 the leading term among those that increase with K, as $K \to \infty$, under the conditions that $a = O(1)$, $b = O(K)$ and $(1-a)K - b = O(1)$ is,

$$\left(\sigma_{u\epsilon}^2\left(\frac{K-2aK-2ab}{N}\right)+\sigma_u^2\sigma_\epsilon^2\frac{K}{N}\right)\left(\frac{f'Pf}{N}\right)^{-2}H^2 = \sigma_{u\epsilon}^2\left(\frac{K-2aK-2ab}{N}\right)+\sigma_u^2\sigma_\epsilon^2\frac{K}{N} + O\left(\left(\frac{K}{N}\right)\frac{f'(I-P)f}{N}\right)$$

This follows since under the condition $(1-a)K - b = O(1)$, we have that the term W_1 in the expansion in Proposition 4, satisfies,

$$W_1 = \frac{u'P\epsilon}{\sqrt{N}} - \left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2\sqrt{N}}+\frac{b}{\sqrt{N}}\right)\sigma_{u\epsilon} = O\left(\left(\frac{K}{N}\right)^{1/2}\right)$$

Then the expression given in the result follows using (14) and collecting leading terms among those that increase with K and those that decrease with K. **Q.E.D.**

Proof of Proposition 3: This result follows directly from the result in Proposition 5. This follows even though we now only use Assumption 3 and not the stronger condition used in Proposition 5, because we are interested in dominant terms among those that increase with K and decrease with K and these are provided in the expression in Proposition 5. Now, use similar arguments to those in the proofs of Propositions 1 and 2, and the fact that by Assumption 3,

$$\begin{aligned}\frac{f'(I-P)\bar{P}(I-P)f}{N} &\leq \sup P_{ii}\frac{f'(I-P)f}{N}\\ &= o\left(\frac{f'(I-P)f}{N}\right)\end{aligned}$$

by Assumption 3 (note that the term $\bar{P}$ is defined in the proof of Proposition 5). Thus, the leading term among those that decrease with K is the term that results from the in (14),

$$\sigma_\epsilon^2\frac{f'(I-P)f}{N}$$

while, using the result in (13) the leading term among those that are increasing in K is given by,

$$(\sigma_{u\epsilon}^2 + \sigma_{u\epsilon}^2)\frac{K}{N}\left(\frac{f'Pf}{N}\right)^{-2} H^2 = (\sigma_{u\epsilon}^2 + \sigma_{u\epsilon}^2)\frac{K}{N} + O\left(\frac{K}{N}\frac{f'(I-P)f}{N}\right)$$

with the second term on the right being of smaller order than the first, using Assumption 2. **Q.E.D.**

Proof of Proposition 4: Denote $\bar{f} = Pf$, $\bar{H} = \bar{f}'\bar{f}/N$ and $\bar{h} = \bar{f}'\epsilon/\sqrt{N}$ and write the estimator as,

$$\begin{aligned}\sqrt{N}(\hat{\gamma} - \gamma) &= \bar{H}^{-1}(\hat{h} + (\bar{H} - \hat{H})\bar{H}^{-1}\bar{h} \\ &\quad +(\bar{H} - \hat{H})\bar{H}^{-1}(\hat{h} - h) + (\bar{H} - \hat{H})^2\bar{H}^{-2}\bar{h} + ...\end{aligned} \tag{15}$$

where,

$$\hat{h} = (\frac{Y'P\epsilon}{\sqrt{N}} - (a\theta + \frac{b}{N})\frac{Y'\epsilon}{\sqrt{N}})$$

and,

$$\hat{H} = (\frac{Y'PY}{N} - (a\theta + \frac{b}{N})\frac{Y'Y}{N})$$

First we deal with $\hat{h}$. Note that using Fujikoshi (1977) the term, θ has the expansion,

$$\theta = \frac{\epsilon'(P - \bar{F})\epsilon}{\sigma_\epsilon^2 N} - R_{N,K}$$

where,

$$\begin{aligned}R_{N,K} &= \frac{1}{N^{3/2}}\left\{\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^4}\sqrt{N}(\frac{\epsilon'\epsilon}{N} - \sigma_\epsilon^2) - 2\frac{\epsilon'(P-\bar{F})v}{\sigma_\epsilon^2\sigma_v}\left(\frac{\bar{f}'\bar{f}}{N}\right)^{-1}\frac{\bar{f}'\epsilon}{\sqrt{N}}\right\} + O\left(\frac{K}{N^2}\right) \\ &= R_{N,K}^1 + O\left(\frac{K}{N^2}\right)\end{aligned}$$

with $\bar{F} = \bar{f}(\ \bar{f}'\ \bar{f})^{-1}\ \bar{f}'$. Throughout this proof, which is done under the assumptions of Proposition 4, with fixed K, we nevertheless indicate the how the order of remainder terms will depend on K when K grows. Thus for instance the remainder term in the last expression is $O(N^{-2})$ when K is fixed. Use will also be made of the simple facts that,

$$\frac{u'\epsilon}{N} = \sigma_{u\epsilon} + \left(\frac{u'\epsilon}{N} - \sigma_{u\epsilon}\right)$$

$$\frac{u'u}{N} = \sigma_u^2 + \left(\frac{u'u}{N} - \sigma_{uu}\right)$$

and,

$$\frac{\epsilon'\epsilon}{N} = \sigma_\epsilon^2 + \left(\frac{\epsilon'\epsilon}{N} - \sigma_{\epsilon\epsilon}\right)$$

with the second term on the right hand side of each expression being $O(N^{-1/2})$ by the Lindberg-Levy Central Limit Theorem. Given this, for the $\hat{h}$ term the relevant expansion is,

$$\hat{h} = \bar{h} + W_1 + W_2 + o(\frac{K}{N})$$

where,

$$\begin{aligned} W_1 &= \frac{u'P\epsilon}{\sqrt{N}} - \left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2\sqrt{N}} + \frac{b}{\sqrt{N}}\right)\sigma_{u\epsilon} = O\left(\frac{K}{\sqrt{N}}\right) \\ W_2 &= -\left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2 N} + \frac{b}{N}\right)\left\{\sqrt{N}\left(\frac{u'\epsilon}{N} - \sigma_{u\epsilon}\right) + h\right\} \\ &\quad -a\sqrt{N}R^1_{N,K}\sigma_{u\epsilon} = O\left(\frac{K}{N}\right) \end{aligned}$$

where the last two expressions are $O(N^{-1/2})$ and $O(N^{-1})$ respectively when K is fixed. Also the remainder term is $O(N^{-1})$ when K is fixed. Also we have,

$$\hat{H} = \bar{H} + W_3 + W_4 + o\left(\frac{K}{N}\right)$$

where,

$$\begin{aligned} W_3 &= 2\left(\frac{\bar{f}'u}{N}\right) = O(N^{-1/2}) \\ W_4 &= \frac{u'Pu}{N} - \left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2 N} + \frac{b}{N}\right)(H + \sigma_u^2) = O\left(\frac{K}{N}\right) \end{aligned}$$

where again the last term is $O(N^{-1})$ under the assumptions of the result.

Using (15) and again grouping terms that are $o(K/N)$ (with K fixed) in the remainder we can write,

$$\sqrt{N}(\hat{\gamma} - \gamma) = \bar{H}^{-1}(\sum_{j=1}^{7} T_j) + o(\frac{K}{N})$$

where,

$$
\begin{aligned}
T_1 &= \bar{h} \\
T_2 &= W_1 = O\left(\frac{K}{\sqrt{N}}\right) \\
T_3 &= -W_3\bar{H}^{-1}\bar{h} = O(N^{-1/2}) \\
T_4 &= W_2 = O\left(\frac{K}{N}\right) \\
T_5 &= -W_4\bar{H}^{-1}\bar{h} = O\left(\frac{K}{N}\right) \\
T_6 &= -W_3\bar{H}^{-1}W_1 = O\left(\frac{K}{N}\right) \\
T_7 &= W_3^2 H^{-2}\bar{h} = O(N^{-1})
\end{aligned}
$$

Then the relevant terms in the MSE are those in the square of this expression that are of the appropriate orders, (again remembering that K is held fixed) and terms up to and including those that are $O(N^{-1})$ will be retained. Also note that since we are conditioning on the exogenous variables the term $\bar{H}^{-1}$ will appear squared in the final result but can be ignored until the final result is presented. Thus we calculate,

$$
\begin{aligned}
E\left[(\sum_{j=1}^{7} T_j)^2\right] &= E(T_1^2) + 2E(T_2T_1) + 2E(T_3T_1) + E(T_2^2) + E(T_3^2) + 2E(T_2T_3) \quad (16) \\
&\quad +2E(T_4T_1) + 2E(T_5T_1) + 2E(T_6T_1) + 2E(T_7T_1) + ...
\end{aligned}
$$

where terms that are $o(N^{-1})$ are omitted from the calculations. We calculate each of the terms in this expression. First,

$$E(T_1^2) = E(\frac{\bar{f}'\epsilon\epsilon'\bar{f}}{N}) = \sigma_\epsilon^2\bar{H}$$

using the fact that,

$$E(\frac{\epsilon\epsilon'}{N}) = \sigma_\epsilon^2 I$$

which follows by Assumption 1. Next, in order to calculate the remaining terms we make repeated use of the fact that for a pair of double arrays, say $\{S_{ij}^1\}$ and $\{S_{ij}^2\}$,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}^1 \sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}^2 = \sum_{i=1}^{N} S_{ii}^1 S_{ii}^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} S_{ii}^1 S_{jj}^2 \quad (17)$$

$$+\sum_{i=1}^{N}\sum_{j\neq 1}^{N} S_{ij}^1 S_{ij}^2 + \sum_{i=1}^{N}\sum_{j\neq 1}^{N} S_{ij}^1 S_{ji}^2 + ...$$

where the omitted terms will not be needed for our calculations because they will have zero expectations in our context. This can be used to calculate the individual terms in (16) the first of which is,

$$\begin{aligned} E(T_2^2) &= E\left(\frac{(u'P\epsilon)^2}{N}\right) + a^2\sigma_{u\epsilon}^2 E\left(\frac{(\epsilon'(P-\bar{F})\epsilon)^2}{\sigma_\epsilon^4 N}\right) + \sigma_{u\epsilon}^2\frac{b^2}{N} \\ &\quad -2a\frac{\sigma_{u\epsilon}}{\sigma_\epsilon^2 N}E\left(u'P\epsilon\epsilon'(P-\bar{F})\epsilon\right) - 2\sigma_{u\epsilon}\frac{b}{N}E(u'P\epsilon) \\ &\quad +2\sigma_{u\epsilon}^2\frac{b}{\sigma_\epsilon^2 N}E(\epsilon'(P-\bar{F})\epsilon) \end{aligned}$$

Using (17) with , $S_{ij}^1 = S_{ij}^2 = u_i\epsilon_j P_{ij}$ along with Assumption 1 we get,

$$\begin{aligned} E\left(\frac{(u'P\epsilon)^2}{N}\right) &= \frac{1}{N}E\left(\sum_{i=1}^{N} u_i^2\epsilon_i^2 P_{ii}^2\right) + \frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i\epsilon_i P_{ii} u_j\epsilon_j P_{jj}\right) \\ &\quad +\frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq 1}^{N} u_i\epsilon_j P_{ij} u_i\epsilon_j P_{ij}\right) + \frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq 1}^{N} u_i\epsilon_j P_{ij} u_j\epsilon_i P_{ji}\right) \end{aligned}$$

Now using Assumptions 1 and 3 and Lemma A.1(i) we have that,

$$\begin{aligned} \frac{1}{N}E\left(\sum_{i=1}^{N} u_i^2\epsilon_i^2 P_{ii}^2\right) &= o(\frac{K}{N}) \\ \frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq i}^{N} u_i\epsilon_i P_{ii} u_j\epsilon_j P_{jj}\right) &= \sigma_{u\epsilon}^2\frac{K^2}{N} + o(\frac{K}{N}) \\ \frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq 1}^{N} u_i\epsilon_j P_{ij} u_i\epsilon_j P_{ij}\right) &= \sigma_u^2\sigma_\epsilon^2\frac{K}{N} + o(\frac{K}{N}) \\ \frac{1}{N}E\left(\sum_{i=1}^{N}\sum_{j\neq 1}^{N} u_i\epsilon_j P_{ij} u_j\epsilon_i P_{ji}\right) &= \sigma_{u\epsilon}^2\frac{K}{N} + o(\frac{K}{N}) \end{aligned}$$

so combined we have that,

$$E\left(\frac{(u'P\epsilon)^2}{N}\right) = \sigma_{u\epsilon}^2\frac{K^2}{N} + (\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2)\frac{K}{N} + o(\frac{K}{N})$$

Next, using similar arguments the fact that $\bar{F}P = \bar{F}$ and Lemma A.2, we have

$$a^2\sigma_{u\epsilon}^2 E\left(\frac{(\epsilon'(P-\bar{F})\epsilon)^2}{\sigma_\epsilon^4 N}\right) = a^2\sigma_{u\epsilon}^2\left(\frac{K^2}{N}-\frac{1}{N}\right)+o\left(\frac{K}{N}\right)$$

$$2a\frac{\sigma_{u\epsilon}}{\sigma_\epsilon^2 N}E\left(u'P\epsilon\epsilon'(P-\bar{F})\epsilon\right) = 2a\sigma_{u\epsilon}^2\left(\frac{K^2}{N}+\frac{K}{N}-\frac{2}{N}\right)+o\left(\frac{K}{N}\right)$$

$$2\sigma_{u\epsilon}\frac{b}{N}E(u'P\epsilon) = 2b\sigma_{u\epsilon}^2\left(\frac{K}{N}\right)$$

and,

$$2a\sigma_{u\epsilon}^2\frac{b}{\sigma_\epsilon^2 N}E(\epsilon'(P-\bar{F})\epsilon) = 2ab\sigma_{u\epsilon}^2\left(\frac{K}{N}-\frac{1}{N}\right)$$

Therefore,

$$\begin{aligned}E(T_2^2) &= \sigma_{u\epsilon}^2\left(\frac{[(1-a)K-b]^2}{N}+\frac{K-2aK-2ab}{N}\right)\\ &\quad +\sigma_{u\epsilon}^2\left(\frac{4a-a^2}{N}\right)+\sigma_u^2\sigma_\epsilon^2\frac{K}{N}\end{aligned}$$

Assumption 1(iii) implies that $E(T_2T_1) = E(T_3T_1) = 0$. Next,

$$\begin{aligned}E(T_3^2) &= E(T_7T_1) = 4E\left(\left[\frac{\bar{f}'u}{N}\bar{H}^{-1}\bar{h}\right]^2\right)\\ &= \frac{4}{N}(\sigma_u^2\sigma_\epsilon^2+2\sigma_{u\epsilon}^2)+o(\frac{1}{N})\end{aligned}$$

using Assumptions 1 and 2 and Lemmas A.1 and A.2. The next result concerns $E(T_4T_1)$. Using Assumption 1(iii) which implies that third moments are zero we have that,

$$\begin{aligned}E(T_4T_1) &= -E\left\{\left\{a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2 N}+\frac{b}{N}\right\}h\bar{h}\right\}\\ &= -\left\{a(\sigma_\epsilon^2\frac{K}{N}+2\sigma_\epsilon^2\frac{1}{N}-3\sigma_\epsilon^2\frac{1}{N})+\sigma_\epsilon^2\frac{b}{N}\right\}\bar{H}\end{aligned}$$

Next,

$$\begin{aligned}E(T_6T_1) &= E(T_2T_3) = E\left[-2\left\{\frac{\bar{f}'u}{N}\right\}\bar{H}^{-1}\bar{h}\left\{\frac{u'P\epsilon}{\sqrt{N}}-\left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2\sqrt{N}}+\frac{b}{\sqrt{N}}\right)\sigma_{u\epsilon}\right\}\right]\\ &= -2E\left(\frac{\bar{f}'u}{N}\bar{H}^{-1}\bar{h}\frac{u'P\epsilon}{\sqrt{N}}\right)+2E\left(a\sigma_{u\epsilon}\frac{\bar{f}'u}{N}\bar{H}^{-1}\bar{h}\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2\sqrt{N}}\right)\\ &\quad +2E\left(\frac{\bar{f}'u}{N}\bar{H}^{-1}\bar{h}\frac{b}{\sqrt{N}}\sigma_{u\epsilon}\right)\end{aligned}$$

$$= -2\left(\sigma_{u\epsilon}^2\frac{K}{N} + \left(\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2\right)\frac{1}{N} - a\sigma_{u\epsilon}^2\frac{K}{N} - 2a\sigma_{u\epsilon}^2\frac{1}{N} + 3a\sigma_{u\epsilon}^2\frac{1}{N} - 2\sigma_{u\epsilon}^2\frac{b}{N}\right)$$
$$+o(\frac{1}{N})$$

Finally we have,

$$\begin{aligned}
E(T_5T_1) &= E\left[-\left\{\frac{u'Pu}{N} - \left(a\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2 N} + \frac{b}{N}\right)(H+\sigma_u^2)\right\}\bar{H}^{-1}\bar{h}^2\right] \\
&= -E\left(\frac{u'Pu}{N}\bar{H}^{-1}\bar{h}^2\right) + aE\left(\frac{\epsilon'(P-\bar{F})\epsilon}{\sigma_\epsilon^2 N}(H+\sigma_u^2)\bar{H}^{-1}\bar{h}^2\right) \\
&\quad +E\left(\frac{b}{N}(H+\sigma_u^2)\bar{H}^{-1}\bar{h}^2\right) \\
&= -(\sigma_u^2\sigma_\epsilon^2\frac{K}{N} + 2\sigma_{u\epsilon}^2\frac{1}{N}) + a(H+\sigma_u^2)(\sigma_\epsilon^2\frac{K}{N} + 2\sigma_\epsilon^2\frac{1}{N} - 3\sigma_\epsilon^2\frac{1}{N}) \\
&\quad +(H+\sigma_u^2)\sigma_\epsilon^2\frac{b}{N} + o(\frac{1}{N})
\end{aligned}$$

Substituting each of these into (16) we get the expression given in the result. **Q.E.D.**

Proof of Proposition 5: Normalized JIVE can be written as,

$$\sqrt{N}(\hat{\gamma}-\gamma) = \left(\frac{Y'C'Y}{N}\right)^{-1}\frac{Y'C'\epsilon}{\sqrt{N}}$$

where,

$$C = P - \bar{P}(I-P)$$

with $\bar{P} = \tilde{P}(I-\tilde{P})^{-1}$ with $\tilde{P}$ being a diagonal matrix with element P_{ii} on the ith diagonal. Use the following notation. Let, $\tilde{f} = Cf$, $\tilde{H} = \tilde{f}'\bar{f}/N$ and $\tilde{h} = \tilde{f}'\epsilon/\sqrt{N}$ and write the estimator as,

$$\begin{aligned}
\sqrt{N}(\hat{\gamma}-\gamma) &= \tilde{H}^{-1}(\hat{h} + (\tilde{H}-\hat{H})\tilde{H}^{-1}\tilde{h} \\
&\quad +(\tilde{H}-\hat{H})\tilde{H}^{-1}(\hat{h}-\tilde{h}) + (\tilde{H}-\hat{H})\tilde{H}^{-2}\tilde{h} + ...
\end{aligned} \tag{18}$$

In this expression we can write,

$$\hat{h} = \frac{Y'C'\epsilon}{\sqrt{N}} = \tilde{h} + W_1$$

where,

$$W_1 = \frac{u'C'\epsilon}{\sqrt{N}} = O\left(\sqrt{\frac{K}{N}}\right)$$

follows from the fact that the diagonals of C are zero and using similar arguments to those used in (17). Also,

$$\hat{H} = \frac{Y'C'Y}{N} = \tilde{H} + W_2 + W_3$$

where,

$$\begin{aligned} W_2 &= \left(\frac{u'C'f}{N} + \frac{f'C'u}{N}\right) = O(N^{-1/2}) \\ W_3 &= \frac{u'C'u}{N} = O\left(\frac{\sqrt{K}}{N}\right) \end{aligned}$$

The order for W_3 follows similarly to that of W_1. To show the order for W_2 we use the following result that will also prove useful below:

$$\frac{f'D'f}{N} = \frac{f'Pf}{N} + O(\sup_i P_{ii}) = \frac{f'Pf}{N} + o(N^{-1}) \tag{19}$$

which holds when D is any one of the matrices C, C', $C'C$, CC, CCC, $C'C'C'$, $C'C'C$ and CCC'. This can be shown in each case using the definition of C the Cauchy Schwarz inequality, repeated use of the inequality,

$$\frac{b'Ab}{N} \leq \frac{b'b}{N} \sup_x \frac{x'Ax}{x'x} = \frac{b'b}{N}\lambda_{\max}(A)$$

where, $\lambda_{\max}(.)$ is the largest eigenvalue of its argument, Assumption 2 and Assumption 3 which implies that the largest element of the diagonal of $\bar{P}$ is $o(1)$. Similarly one can show that,

$$\frac{1}{\sqrt{N}}|\ddot{f}_i| \leq \left(\frac{f'f}{N}\right)^{1/2} \left(\sup_i P_{ii}\right)^{1/2} \tag{20}$$

for $\ddot{f}_i$ equal to the ith element of any of the vectors $\tilde{f}$, $\check{f} = C'f$ or $\bar{f} = Pf$ (the last of which was shown in Lemma A.2).

As with the proof of Proposition 4, although K is fixed, we have nevertheless indicated how the order of remainder terms will depend on K when K grows.

Using (18) and again grouping terms that are $o(K/N)$ (with K fixed) in the remainder we can write,

$$\sqrt{N}(\hat{\gamma} - \gamma) = \tilde{H}^{-1}(\sum_{j=1}^{6} T_j) + o(\frac{K}{N})$$

where,

$$\begin{aligned}
T_1 &= \bar{h} \\
T_2 &= W_1 = O\left(\sqrt{\frac{K}{N}}\right) \\
T_3 &= -W_2\tilde{H}^{-1}\tilde{h} = O(N^{-1/2}) \\
T_4 &= -W_3\tilde{H}^{-1}\tilde{h} = O\left(\frac{\sqrt{K}}{N}\right) \\
T_5 &= -W_2\tilde{H}^{-1}W_1 = O\left(\frac{\sqrt{K}}{N}\right) \\
T_6 &= W_2^2\tilde{H}^{-2}\tilde{h} = O(N^{-1})
\end{aligned}$$

Then the relevant terms in the MSE are those in the square of this expression that are of the appropriate orders, (again remembering that K is held fixed) and terms up to and including those that are $O(N^{-1})$ will be retained. Also note that since we are conditioning on the exogenous variables the term $\tilde{H}^{-1}$ will appear squared in the final result but can be ignored until the final result is presented. Thus we calculate,

$$\begin{aligned}
E\left[(\sum_{j=1}^{6} T_j)^2\right] &= E(T_1^2) + 2E(T_2T_1) + 2E(T_3T_1) + E(T_2^2) + E(T_3^2) \qquad (21) \\
&\quad +2E(T_2T_3) + 2E(T_4T_1) + 2E(T_5T_1) + 2E(T_6T_1) + ... \qquad (22)
\end{aligned}$$

where terms that are $o(N^{-1})$ are omitted from the calculations. We calculate each of the terms in this expression. First,

$$E(T_1^2) = E(\frac{\tilde{f}'\epsilon\epsilon'\tilde{f}}{N}) = \sigma_\epsilon^2\frac{\tilde{f}'\tilde{f}}{N}$$

using the fact that,

$$E(\frac{\epsilon\epsilon'}{N}) = \sigma_\epsilon^2 I$$

which follows by Assumption 1. Next, using (17) we have that,

$$\begin{aligned} E(T_2^2) &= E\left(\frac{(u'C'\epsilon)^2}{N}\right) = E\left(\frac{\sum_{i\neq j} u_i^2\epsilon_j^2 C_{ij}^2}{N}\right) + E\left(\frac{\sum_{i\neq j} u_i\epsilon_i u_j\epsilon_j C_{ij}C_{ji}}{N}\right) \\ &= \frac{\sigma_u^2\sigma_\epsilon^2\sum_{i\neq j} C_{ij}^2}{N} + \frac{\sigma_{u\epsilon}^2\sum_{i\neq j} C_{ij}C_{ji}}{N} \\ &= (\sigma_u^2\sigma_\epsilon^2 + \sigma_{u\epsilon}^2)\frac{K}{N} + o(\sup_i P_{ii}\frac{K}{N}) \end{aligned}$$

Note that in the last line we have used Lemma A.1 and the fact that,

$$\begin{aligned} |\sum_{i\neq j}(C_{ij}^2 - P_{ij}^2)| &\leq \sum_{i\neq j} P_{ij}^2|(1-P_{ii})^{-2} - 1| \\ &\leq \sup_i |(1-P_{ii})^{-2} - 1| \sum_{i\neq j} P_{ij}^2 \\ &= O(\sup_i P_{ii})(K + O(\sup_i P_{ii})) \end{aligned}$$

which follows from Assumption 3, and Lemma A.1. Similarly,

$$\begin{aligned} |\sum_{i\neq j}(C_{ij}C_{ji} - P_{ij}P_{ji})| &= \sum_{i\neq j} P_{ij}^2|(1-P_{ii})(1-P_{jj}) - 1| \\ &\leq \sup_i |(1-P_{ii})^{-2} - 1| \sum_{i\neq j} P_{ij}^2 \\ &= O(\sup_i P_{ii})(K + O(\sup_i P_{ii})). \end{aligned}$$

Next we have that, by Assumption 1(iii), $E(T_2T_1) = E(T_3T_1) = 0$. Next,

$$\begin{aligned} E(T_3^2) &= E(T_6T_1) = E\left(\left[\frac{\tilde{f}'u + u'\check{f}}{N}\tilde{H}^{-1}\tilde{h}\right]^2\right) \\ &= E\left((\frac{\tilde{f}'u}{N})^2\tilde{H}^{-2}\tilde{h}^2\right) + E\left((\frac{\check{f}'u}{N})^2\tilde{H}^{-2}\tilde{h}^2\right) + 2E\left(\frac{\tilde{f}'uu'\check{f}}{N}\tilde{H}^{-2}\tilde{h}^2\right) \\ &= \frac{4}{N}(\sigma_u^2\sigma_\epsilon^2 + 2\sigma_{u\epsilon}^2) + o(\frac{1}{N}) \end{aligned}$$

using Assumptions 1 and 2 and the results in (19) and (20). Next,

$$E(T_5T_1) = E(T_2T_3) = E\left[-\left\{\frac{\tilde{f}'u + u'\check{f}}{N}\right\}\tilde{H}^{-1}\tilde{h}\left\{\frac{u'C'\epsilon}{\sqrt{N}}\right\}\right]$$

$$
\begin{aligned}
&= -E\left(\frac{\tilde{f}'u}{N}\tilde{H}^{-1}\tilde{h}\frac{u'C'\epsilon}{\sqrt{N}}\right) - E\left(\frac{\tilde{f}'u}{N}\tilde{H}^{-1}\tilde{h}\frac{u'C'\epsilon}{\sqrt{N}}\right) \\
&= -\frac{2}{N}\left(\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2\right) + o(\frac{1}{N})
\end{aligned}
$$

again using Assumptions 1 and 2 and the results in (19) and (20). Finally we have, using similar arguments,

$$
\begin{aligned}
E(T_4T_1) &= -E\left[(\frac{u'C'u}{N})\tilde{H}^{-1}\tilde{h}^2\right] \\
&= -2\sigma_{u\epsilon}^2\frac{1}{N} + o(\frac{1}{N})
\end{aligned}
$$

Substituting each of these into (21) and multiplying by $\tilde{H}^{-2}$ we get the expression,

$$
\begin{aligned}
&\sigma_\epsilon^2\frac{f'C'Cf}{N}\left(\frac{f'C'f}{N}\right)^{-2} + (\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2)\frac{K}{N}\left(\frac{f'C'f}{N}\right)^{-2} \\
&+ \left(12\sigma_{u\epsilon}^2 + 4\sigma_u^2\sigma_\epsilon^2\right)\frac{1}{N}\left(\frac{f'C'f}{N}\right)^{-2}
\end{aligned} \tag{23}
$$

Now consider the following,

$$
\frac{f'C'f}{N} = \frac{f'Pf}{N} - \frac{f'\bar{P}(I-P)'f}{N}
$$

with,

$$
\begin{aligned}
\left|\frac{f'\bar{P}(I-P)f}{N}\right| &\leq \left(\frac{f'\bar{P}^2f}{N}\right)^{1/2}\left(\frac{f'(I-P)f}{N}\right)^{1/2} \\
&\leq \sup_i\frac{P_{ii}}{1-P_{ii}}\left(\frac{f'f}{N}\right)^{1/2}\left(\frac{f'(I-P)f}{N}\right)^{1/2} \\
&= O(N^{-1})
\end{aligned} \tag{24}
$$

given the additional condition that $\sup_i P_{ii} = O(N^{-1})$. Next,

$$
\begin{aligned}
\frac{f'C'Cf}{N} &= \frac{f'Pf}{N} - 2\frac{f'\bar{P}(I-P)f}{N} && (25) \\
&\quad +2\frac{f'(I-P)\bar{P}(I-P)f}{N} + \frac{f'(I-P)\bar{P}^2(I-P)f}{N} && (26)
\end{aligned}
$$

where the term,

$$
\begin{aligned}
\frac{f'(I-P)\bar{P}(I-P)f}{N} &\leq \sup_i\frac{P_{ii}}{1-P_{ii}}\left(\frac{f'(I-P)f}{N}\right) \\
&= O(N^{-1})
\end{aligned}
$$

and,

$$\begin{aligned}\frac{f'(I-P)\bar{P}^2(I-P)f}{N} &\leq \sup_i \left(\frac{P_{ii}}{1-P_{ii}}\right)^2 \left(\frac{f'(I-P)f}{N}\right) \\ &= O(N^{-2})\end{aligned}$$

using similar arguments. Next use the expansion,

$$\left(\frac{f'C'f}{N}\right)^{-1} = \left(\frac{f'Pf}{N}\right)^{-1} + \frac{f'\bar{P}(I-P')f}{N}\left(\frac{f'Pf}{N}\right)^{-2} + O\left(N^{-2}\right)$$

substitute this along with expansion (25) into the expression (23) and omit terms that are $o(N^{-1})$ and we obtain,

$$\sigma_\epsilon^2 \frac{f'C'Cf}{N}\left(\frac{f'C'f}{N}\right)^{-2} = \sigma_\epsilon^2\left(\frac{f'Pf}{N}\right)^{-1} + 2\sigma_\epsilon^2\left(\frac{f'(I-P)\bar{P}(I-P)f}{N}\right)\left(\frac{f'Pf}{N}\right)^{-2}$$

Similarly we have that,

$$(\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2)\frac{K}{N}\left(\frac{f'C'f}{N}\right)^{-2} = (\sigma_{u\epsilon}^2 + \sigma_u^2\sigma_\epsilon^2)\frac{K}{N}\left(\frac{f'Pf}{N}\right)^{-2} + O\left(\frac{K}{N^2}\right)$$

and,

$$\left(12\sigma_{u\epsilon}^2 + 4\sigma_u^2\sigma_\epsilon^2\right)\frac{1}{N}\left(\frac{f'C'f}{N}\right)^{-2} = \left(12\sigma_{u\epsilon}^2 + 4\sigma_u^2\sigma_\epsilon^2\right)\frac{1}{N}\left(\frac{f'Pf}{N}\right)^{-2} + O(N^{-2})$$

so that we end up with the expression given in the result. **Q.E.D.**

Proof of Proposition 6: Notice that since the choice of K is unaffected by the removal of constants from $\hat{S}_L(K)$ we can assume without loss of generality that $\hat{S}_L(K)$ has been constructed using, $\tilde{R}(K) = \hat{R}(K) - \bar{\sigma}_u^2$. Using Lemma A.3, the proof of which is given below, we need only to show that,

$$\sup_K \frac{|\hat{S}_L(K) - S_L(K)|}{S_L(K)} = o_p(1).$$

To show this, note that,

$$\begin{aligned}&\sup_K \frac{|\hat{S}_L(K) - S_L(K)|}{S_L(K)} \\ &\leq \sup_K \frac{(|\hat{\sigma}_v^2 - \sigma_v^2)| + |\hat{\sigma}_u^2 - \sigma_u^2|)K/N}{\sigma_v^2 K/N} + \sup_K \frac{|\tilde{R}(K) - R(K)|}{R(K)}\end{aligned}$$

The first term is $o_p(1)$ by Assumption 4 (i) and (ii) while the second term is $o_p(1)$ given the result of Lemma A.4, which follows from Assumption 5. **Q.E.D.**

Proof of Proposition 7: As was the case with the Proof of Proposition 6, we can assume without loss of generality that $\hat{S}_2(K)$ has been constructed using $\tilde{R}(K) = \hat{R}(K) - \bar{\sigma}_u^2$. Similar to the proof of Proposition 6, we need only to show that,

$$\sup_K \frac{|\hat{S}_2(K) - S_2(K)|}{S_2(K)} = o_p(1)$$

where $\hat{S}_2(K)$ and $S_2(K)$ are the estimated and actual approximate mean squared error criteria for 2SLS. To show this, note that,

$$\sup_K \frac{|\hat{S}_2(K) - S_2(K)|}{S_2(K)} \leq \sup_K \frac{|\hat{\sigma}_{u\epsilon}^2 - \sigma_{u\epsilon}^2| K^2/N}{S_2(K)}$$

$$+ \sup_K \frac{|\hat{\sigma}_\epsilon^2(\tilde{R}(K) - R(K)) + (\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)R(K) - (\hat{\sigma}_u^2 - \sigma_u^2)K/N|}{S_2(K)}$$

$$\leq \frac{|\hat{\sigma}_{u\epsilon}^2 - \sigma_{u\epsilon}^2|}{\sigma_{u\epsilon}^2} + \hat{\sigma}_u^2 \sup_K \frac{|\tilde{R}(K) - R(K)|}{S_2(K)}$$

$$+ |\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2| \sup_K \frac{R(K)}{S_2(K)} + |\hat{\sigma}_u^2 - \sigma_u^2| \sup_K \frac{K/N}{S_2(K)}$$

We consider the four terms in the last expression separately. The first term is $o_p(1)$ by Assumption 4(iii). Notice that, using the definition of $R(K)$ we have that,

$$\frac{R(K)}{S_2(K)} \leq \frac{\sigma_u^2}{\sigma_{u\epsilon}} + \frac{1}{\sigma_\epsilon^2} \tag{27}$$

using the fact that $K^2/N \geq K/N$. Using these facts we have that for the second term,

$$\hat{\sigma}_\epsilon^2 \sup_K \frac{|\tilde{R}(K) - R(K)|}{S_2(K)}$$

$$\leq \hat{\sigma}_\epsilon^2 \sup_K \frac{|\tilde{R}(K) - R(K)|}{R(K)} \sup_K \frac{R(K)}{\sigma_{u\epsilon}^2 K/N + \sigma_\epsilon^2 f'(I-P)f/N}$$

which is $o_p(1)$ using Assumption 4(i), (27) and Lemma A.4 which follows from Assumption 5. For the third term we have similarly that,

$$|\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2| \sup_K \frac{R(K)}{S_2(K)} \leq \sup_K \frac{|\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2| R(K)}{\sigma_{u\epsilon}^2 K/N + \sigma_\epsilon^2 f'(I-P)f/N}$$

$$= o_p(1)$$

again using (27) and Assumption 5. Finally the last term is easily seen to be $o_p(1)$ using (ii) and the fact that

$$\sup_K \frac{K/N}{\sigma_{u\epsilon}^2 K^2/N + \sigma_\epsilon^2 f'(I-P)f/N} \leq \sup_K \frac{1}{\sigma_{u\epsilon}^2 K}$$

which must be bounded since $K \geq 1$ (otherwise the order condition would not be satisfied!). **Q.E.D.**

Proof of Proposition 8: As in the two previous proofs, we can assume without loss of generality that $\hat{S}_J(K)$ has been constructed using, $\tilde{R}(K) = \hat{R}(K) - \bar{\sigma}_u^2$.Again we must show that

$$\sup_K \frac{|\hat{S}_J(K) - S_J(K)|}{S_J(K)} = o_p(1).$$

To show this, note that,

$$\begin{aligned} \sup_K \frac{|\hat{S}_J(K) - S_J(K)|}{S_J(K)} &\leq \sup_K \frac{(|\hat{\sigma}_{u\epsilon}^2 - \sigma_{u\epsilon}^2)|K/N}{\sigma_{u\epsilon}^2 K/N} \\ &+ \frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2} \sup_K \frac{|\tilde{R}(K) - R(K)|}{R(K)} + \sup_K \frac{|\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2|}{\sigma_\epsilon^2} \end{aligned}$$

The first and third terms are $o_p(1)$ using Assumption 4. The second term is $o_p(1)$ using Lemma A.4, which follows from Assumption 5, and Assumption 4(ii).**Q.E.D.**

Lemma A.1 *Given Assumption 3 the following results hold for the matrix* $P = \Psi(\Psi'\Psi)^{-1}\Psi'$,

(i) $\sum_{i=1}^N P_{ii}^2 = o(K)$,

(ii) $\sum_{i=1}^N \sum_{j\neq i}^N P_{ii}P_{jj} = K^2 - o(K)$,

(iii) $\sum_{i=1}^N \sum_{j\neq i}^N P_{ij}P_{ij} = \sum_{i=1}^N \sum_{j\neq i}^N P_{ji}P_{ij} = K - o(K)$

Proof: To show (i) note that,

$$\sum_{i=1}^N P_{ii}^2 \leq (\sup_i P_{ii}) \sum_{i=1}^N P_{ii} = (\sup_i P_{ii})K = o(1)K = o(K)$$

using the fact that P is a projection matrix and Assumption 3. For (ii) note that,

$$\sum_{i=1}^N \sum_{j\neq i}^N P_{ii}P_{jj} = \sum_{i=1}^N P_{ii} \sum_{j=1}^N P_{jj} - \sum_{i=1}^N P_{ii}^2 = K^2 - o(K)$$

using the fact that P is a projection matrix and result (i). Finally, using the fact that P is a symmetric projection matrix we have that,

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N} P_{ij}P_{ij} = \sum_{i=1}^{N}\sum_{j\neq i}^{N} P_{ji}P_{ij} = trace(P'P) - \sum_{i=1}^{N} P_{ii}^2 = trace(P) - o(K) = K - o(K)$$

Q.E.D.

Lemma A.2 *Given Assumption 3 and either Assumption 2 if* $K \to \infty$ or

$$0 < c < \frac{f'Pf}{N} < c^{-1}$$

with probability 1, uniformly in N, (for some small constant c), the following results hold for the matrix $\bar{F} = Pf(f'Pf)^{-1}f'P$ and the vector $\bar{f} = Pf$,

(i) $\sum_{i=1}^{N} \bar{F}_{ii} = 1$,

(ii) $\sup_i \bar{F}_{ii} = o(1)$

(iii) $\sup_i |\bar{f}_i|/\sqrt{N} = O(\sup_i P_{ii}^{1/2}) = o(1)$

(iv) $\bar{f}'\bar{f}/N = O(1)$

(v) $N^{-1}\sum_{i\neq j} \bar{f}_i\bar{f}_j P_{ij} = \bar{f}'\bar{f}/N + o(1)$

(vi) $N^{-2}\sum_{i\neq j} \bar{f}_j\bar{f}_j P_{ii} = (\bar{f}'\bar{f}/N)(K/N) + o(N^{-1})$.

Proof: (i) This follows from the fact that

$$\sum_{i=1}^{N} \bar{F}_{ii} = trace(Pf(f'Pf)^{-1}f'P) = trace(f'Pf(f'Pf)^{-1}) = 1$$

To show (ii), letting P_i denote the ith column of P (so that $P_i'P_i = P_{ii}$) then,

$$\begin{aligned}
\bar{F}_{ii} &= P_i'f(f'Pf)^{-1}f'P_i \\
&\leq P_{ii}trace(f(f'Pf)^{-1}f') \\
&= P_{ii}trace((\frac{f'Pf}{N})^{-1}\frac{f'f}{N}) \\
&\leq \sup_i P_{ii}O(1)
\end{aligned}$$

where the last line follows from either Assumption 2 or the additional condition in the statement of the result. Then (ii) follows from Assumption3.

(iii) By Cauchy Schwarz,

$$\begin{aligned}\frac{1}{\sqrt{N}}|P_i'f| &\leq (P_i'P_i)^{1/2}\left(\frac{f'f}{N}\right)^{1/2} \\ &\leq \left(\sup_i P_{ii}\right)^{1/2}\left(\frac{f'f}{N}\right)^{1/2} \\ &= o(1)\end{aligned}$$

where the last line follows from Assumptions 3 and 2. To show (iv),

$$\frac{f'Pf}{N} \leq \frac{f'f}{N} = O(1)$$

by Assumption 2. For (v),

$$\frac{\sum_{i\neq j}\bar{f}_i\bar{f}_jP_{ij}}{N} = \frac{\bar{f}'\bar{f}}{N} - \frac{\sum_i \bar{f}_i\bar{f}_iP_{ii}}{N}$$

where,

$$\frac{\sum_i \bar{f}_i\bar{f}_iP_{ii}}{N} \leq \sup_i P_{ii}\frac{\sum_i \bar{f}_i^2}{N} = o(1)$$

by Assumption 3 and result (iv). Finally, (vi) follows for the same reason. **Q.E.D.**

Lemma A.3 *A sufficient condition for a method of selecting K based on,*

$$\hat{K} = arg\min_K \hat{S}_N(K)$$

to be "higher order asymptotically optimal" is the condition that,

$$\sup_K \frac{|\hat{S}_N(K) - S_N(K)|}{S_N(K)} = o_p(1).$$

Proof: By construction,

$$S(\hat{K}) \geq \inf_K S(K). \tag{28}$$

Hence,

$$S(\hat{K}) \geq (1/2)(\inf_K S(K) + S(\hat{K})). \tag{29}$$

Now using (28) and (29),

$$0 < \frac{S(\hat{K}) - \inf_K S(K)}{S(\hat{K})}$$

$$< 2\frac{S(\hat{K}) - \inf_K S(K)}{S(\hat{K}) + \inf_K S(K)} < 2 \sup_K \frac{S(\hat{K}) - S(K)}{S(\hat{K}) + S(K)}$$
$$\leq 2 \sup_K \frac{(S(\hat{K}) - \hat{S}(\hat{K})) - (S(K) - \hat{S}(K))}{S(\hat{K}) + S(K)}$$
$$\leq 2\left(\frac{|S(\hat{K}) - \hat{S}(\hat{K})|}{S(\hat{K})} + \sup_K \frac{|S(K) - \hat{S}(K)|}{S(K)}\right)$$

where the fourth inequality follows from the fact that,

$$\hat{S}(\hat{K}) - \hat{S}(K) \leq 0$$

by the definition of $\hat{K}$. Since the terms in the last expression are $o_p(1)$ by the condition of the Lemma it follows that,

$$\frac{\inf_K S(K)}{S(\hat{K})} \xrightarrow{p} 1$$

and hence by the Slutsky Theorem,

$$\frac{S(\hat{K})}{\inf_K S(K)} \xrightarrow{p} 1$$

and so $\hat{K}$ is asymptotically optimal with respect to $S(K)$. **Q.E.D.**

Lemma A.4: *Given Assumption 5 (i) and (ii) if $\hat{R}(K)$ is the Mallows criteria and Assumption 5 (i) (ii) and (iii) if $\hat{R}(K)$ is the cross validation criteria, we have that,*

$$\sup_K \frac{|\hat{R}(K) - \bar{\sigma}_u^2 - R(K)|}{R(K)} = o_p(1).$$

Proof: For Mallows criteria the result is identical to Theorem 2.1 of Li (1987). For cross validation the proof is adapted from Li (1987). First, we write,

$$\hat{R}(K) - \bar{\sigma}_u^2 - R(K) = \sum_{j=1}^{14} T_j$$

where the T_j differ from those used in other proofs. Throughout, we will let Δ denote a generic large constant. The individual terms are:

$$T_1 = \frac{u'Pu}{N} - \sigma_u^2 \frac{K}{N}$$

$$
\begin{aligned}
T_2 &= \frac{2}{N} f'(I-P)u \\
T_3 &= 2\left(\frac{u'\tilde{P}u}{N} - \frac{u'Pu}{N}\right) \\
T_4 &= \frac{4}{N} f'(I-P)\tilde{P}u \\
T_5 &= -\frac{4}{N} u'P\tilde{P}u \\
T_6 &= -\frac{4}{N} f'(I-P)\tilde{P}Pu \\
T_7 &= \frac{2}{N} u'P\tilde{P}Pu \\
T_8 &= \frac{2}{N} f'(I-P)\tilde{P}(I-P)f \\
T_9 &= \frac{u'\tilde{D}u}{N}
\end{aligned}
$$

where $\tilde{P}$ is a diagonal matrix which has P_{ii}on its ith diagonal and where $\tilde{D}$ is the diagonal matrix with the term $D_{ii} = P_{ii}^2(3-2P_{ii})(1-P_{ii})^{-2}$on the ith diagonal. The terms T_9 through T_{14} are identical to T_4 through T_8 but with $\tilde{P}$ replaced by $\tilde{D}$, 2 replaced by 1 and 4 replaced by 2. These terms are obtained using the fact that,

$$(1-P_{ii})^{-2} = 1 + 2P_{ii} + D_{ii}.$$

The proof proceeds by showing that,

$$\sup_K \frac{|T_j|}{NR(K)} = o_p(1) \tag{30}$$

where the supremum is taken over the relevant index set. Notice that this result holds for $j = 1, 2$, using Theorem 2.1 of Li (1987) given Assumption 5(i) and (ii). Also, note that the results will hold for $j = 10, 11, 12, 13, 14$, if we can show the corresponding results for $j = 4, 5, 6, 7, 8$. Thus we must show the results for $j = 3, 4, 5, 6, 7, 8, 9$. To do this we follow Li (1987) and use the generalized Chebyshev inequality which implies that,

$$P(\sup_K \frac{|T_j|}{NR(K)} > \xi) \leq \sum_K \frac{E(|T_j|^4)}{\xi^4 N^4 R(K)^4} \tag{31}$$

for any $\xi > 0$, where the sum on the right hand side is over the relevant index set. For $j = 3, 4, 6$ this is done by showing that,

$$E(|T_j|^4) \leq \Delta N^2 R(K)^2 \tag{32}$$

which is sufficient to show (31) given that Assumption 5(ii) holds. First, for $j = 3$, we have that (32) follows because, using Theorem 2 of Whittle (1960), Assumption 5(i) and (iii) and the definition of $R(K)$. Using the same result and Assumptions we have for $j = 4$, that,

$$\begin{aligned} E(|T_4|^4) &\leq \Delta(f'(I-P)\tilde{P}\tilde{P}(I-P)f)^2 \\ &\leq \Delta(\sup_K \sup_i P_{ii}^4)(f'(I-P)f)^2 \\ &= o_p(1)N^2R(K)^2 \end{aligned}$$

so that (32) holds for $j = 4$. Similarly for $j = 6$, we have that,

$$\begin{aligned} E(|T_6|^4) &\leq \Delta(f'(I-P)\tilde{P}P\tilde{P}(I-P)f)^2 \\ &\leq \Delta(f'(I-P)\tilde{P}\tilde{P}(I-P)f)^2 = o_p(1)N^2R(K)^2 \end{aligned}$$

where the second inequality follows from the fact that the largest eigenvalue of P is less than 1 due to the fact that P is a projection matrix.

To show (30) for $j = 5$ we first note that,

$$u'\tilde{P}Pu = \sum_{i=1}^{N} u_i^2 P_{ii}^2 + \sum_{i \neq j}^{N} P_{ii}P_{ij}u_iu_j \tag{33}$$

and consider the two terms separately. For the first, note that,

$$\sup_K \frac{|\sum_{i=1}^N u_i^2 P_{ii}^2 - \sigma_u^2 \sum_{i=1}^N P_{ii}^2|}{NR(K)} = o_p(1)$$

can be shown using (31), (32), Theorem 2 of Whittle (1960) and Assumption 5. Then since,

$$\begin{aligned} \sup_K \frac{|\sigma_u^2 \sum_{i=1}^N P_{ii}^2|}{NR(K)} &\leq (\sup_K \sup_i P_{ii}) \sup_K (\frac{|\sigma_u^2 K|}{NR(K)}) \\ &\leq o_p(1) \end{aligned}$$

we have that the first term in (33) satisfies(30). For the second term in (33) we similarly can show that it satisfies (30) by using (31), (32), Theorem 2 of Whittle (1960) and

Assumption 5. To see this note that,

$$\begin{aligned} E(|\sum_{i\neq j}^{N} P_{ii}P_{ij}u_iu_j|^4) &\leq \Delta(\sum_{i\neq j}^{N} P_{ii}^2P_{ij}^2)^2 \\ &\leq \Delta(\sup_K \sup_i P_{ii})^4(\sum_{i\neq j}^{N} P_{ij}^2)^2 \\ &\leq o_p(1)N^2R(K)^2 \end{aligned}$$

using Lemma A.1(iii). Therefore the second term in (33) satisfies (30), so by the triangle inequality (30) is satisfied for T_5. For T_7we can note that, by the fact that,

$$\sup_K \frac{|u'P\tilde{P}Pu|}{NR(K)} \leq (\sup_K \sup_i P_{ii}) \sup_K \frac{|u'Pu|}{NR(K)}$$

we have that (30) is satisfied by Assumption 5(iii) and the fact that

$$\begin{aligned} \sup_K \frac{|u'Pu|}{NR(K)} &\leq \sup_K \frac{|u'Pu - \sigma_u^2 K|}{NR(K)} + \sup_K (\frac{|\sigma_u^2 K|}{NR(K)}) \\ &\leq o_p(1) + 1 \end{aligned}$$

using Theorem 2.1 of Li (1987) and the definition of $R(K)$. Next, (30) is satisfied for T_8 using the fact that,

$$\begin{aligned} f'(I-P)\tilde{P}(I-P)f &\leq (\sup_K \sup_i P_{ii})(f'(I-P)f) \\ &\leq o_p(1)NR(K) \end{aligned}$$

where the second inequality follows from the definition of $R(K)$. Finally, (30) is satisfied for T_9 using similar arguments to those used to deal with the first term of T_5. **Q.E.D.**

REFERENCES

Anderson, T.W. and T. Sawa (1979): "Evaluation of the Distribution Function of the Two Stage Least Squares Estimator", *Econometrica,* 47, 163-182.

Andrews, D.W.K. (1991a): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica,* 59, 817-858.

Andrews, D.W.K. (1991b): "Asymptotic Optimality of Generalized C, Cross-Validation and Generalized Cross-Validation in Regression with Heteroskedastic Errors", *Journal of Econometrics,* 47, 359-377.

Andrews, D.W.K. (1996): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," mimeo.

Angrist, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review,* 80, 313-335.

Angrist, J. and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics,* 106, 979–1014.

Angrist, J.D., G.W. Imbens and A. Krueger (1995): "Jackknife Instrumental Variables Estimation", NBER Technical Working Paper No. 172.

Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica,* 63, 657-681.

Bound, J., D. Jaeger, and R. Baker (1996): "Problems with Instrumental Variables Estimation when the Correlation Between Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association,* 90, 443-450.

Fujikoshi, Y. (1977): "Asymptotic Expansions for the Distributions of Some Multivariate Tests", in *Multivariate Analysis* Vol IV, 55–71.

Li, K.-C. (1987): "Asymptotic Optimality of C_p C, Cross-Validation and Generalized Cross-Validation: Discrete Index Set", *Journal of Econometrics,* 47, 359-377.

Linton, O. (1995): "Second-Order Approximation in a Partially Linear Regression Model," *Econometrica,* 63, 1079-1112.

Maddala, G.S. and J. Jeong (1992), "On the Exact Small Sample Distribution of the Instrumental Variables Estimator", *Econometrica,* 60, 181-183.

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica,* 51, 821-841.

Nagar, A.L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations", *Econometrica,* 27, 575–595.

Nelson, C. and R. Startz (1990): "The Distribution of the Instrumental Variables Estimator and its t-ratio when the Instrument is a Poor One", *Journal of Business, 63, 5125-5140.*

Newey, W.K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models", *Econometrica,* 58, 809-837.

Powell, J.L. and T.M. Stoker (1996), "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics,* 75, 291-316.

Rothenberg, T.J. (1983): "Asymptotic Properties of Some Estimators in Structural Models", in *Studies in Econometrics, Time Series and Multivariate Statistics,* edited by Karlin, S. T Amemiya and L.A. Goodman. New York : Academic Press.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments", *Econometrica, 65, 557-586.*

Table 1.1: Case 1[12]

			Quantiles				
Est.	N	Max K	.10	.50	.90	MAE	Cov. Rate
OLS	100	*	.508	.588	.668	.588	0
2SLS	100	10	-.247	.016	.208	.117	.882
LIML	100	10	-.247	.018	.198	.114	.896
Jackknife IV	100	10	-.325	-.016	.190	.128	.893
2SLS	100	20	-.245	.018	.211	.118	.876
LIML	100	20	-.242	.020	.209	.116	.882
Jackknife IV	100	20	-.314	-.009	.203	.129	.877
2SLSALL	100	20	.137	.282	.410	.282	.220
LIMLALL	100	20	-.315	-.005	.202	.127	.909
OLS	400	*	.548	.587	.628	.588	0
2SLS	400	10	-.116	.004	.101	.057	.898
LIML	400	10	-.114	.006	.101	.057	.903
Jackknife IV	400	10	-.123	-.001	.100	.057	.895
2SLS	400	20	-.112	.001	.103	.056	.897
LIML	400	20	-.111	.004	.106	.056	.895
Jacknife IV	400	20	-.124	-.004	.100	.058	.901
2SLSALL	400	20	-.001	.092	.181	.093	.627
LIMLALL	400	20	-.121	-.002	.102	.057	.902

[12]Note that the entries for OLS, 2SLSALL and LIMLALL are taken from Angrist, Imbens and Krueger (1995, Table 2). The estimators 2SLSALL and LIMLALL use all $K = 20$ instruments in addition to the constant.

Table 1.2: Case 1, Frequencies

	N=100						N=400					
K	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV
1	84.18	48.34	70.06	83.94	44.04	69.70	89.00	48.00	70.94	88.70	44.62	71.32
2	10.58	12.36	11.84	10.44	12.06	11.32	8.76	11.60	11.92	9.12	11.08	11.22
3	3.06	8.84	6.40	3.56	8.08	6.54	1.94	8.08	5.90	1.72	6.66	5.62
4	1.28	6.72	4.10	.96	6.00	3.18	.24	6.66	3.66	.46	5.48	3.46
5	.38	5.26	2.22	.46	4.50	2.66	.06	4.68	2.48	0	4.20	2.06
6	.28	4.40	1.58	.22	4.00	1.44	0	4.54	1.58	0	3.36	1.20
7	.04	4.10	1.16	.12	2.92	1.12	0	4.28	1.40	0	2.82	1.32
8	.08	3.26	1.08	.06	2.72	.94	0	4.00	.88	0	2.88	.92
9	.06	3.36	.80	.10	1.88	.60	0	3.56	.58	0	2.82	.72
10	.06	3.36	.76	.06	2.16	.54	0	4.60	.66	0	2.08	.52
11				0	1.72	.56				0	2.04	.44
12				.02	1.50	.44				0	1.88	.24
13				.02	1.34	.22				0	1.62	.18
14				.02	1.30	.20				0	1.32	.26
15				0	.92	.06				0	1.18	.08
16				0	.98	.14				0	1.40	.08
17				.02	1.24	.08				0	1.04	.06
18				0	.92	.12				0	1.04	.06
19				0	.98	.06				0	1.14	.18
20				0	.74	.08				0	1.34	.06

Table 2.1: Case 2[13]

			Quantiles				
Est.	N	Max K	.10	.50	.90	MAE	Cov. Rate
OLS	100	*	.506	.589	.671	.588	0
2SLS	100	10	-.106	.104	.282	.131	.774
LIML	100	10	-.257	.017	.205	.121	.887
Jackknife IV	100	10	-.377	-.027	.194	.134	.867
2SLS	100	20	-.120	.097	.275	.131	.787
LIML	100	20	-.246	.018	.211	.121	.884
Jackknife IV	100	20	-.344	-.007	.216	.135	.848
2SLSALL	100	20	.134	.284	.414	.284	.231
LIMLALL	100	20	-.305	-.001	.208	.128	.898
OLS	400	*	.549	.589	.629	.589	0
2SLS	400	10	-.091	.023	.118	.058	.877
LIML	400	10	-.114	.004	.102	.057	.901
Jackknife IV	400	10	-.131	-.004	.098	.059	.892
2SLS	400	20	-.094	.022	.120	.059	.868
LIML	400	20	-.114	.006	.107	.058	.888
Jackknife IV	400	20	-.132	-.004	.103	.062	.881
2SLSALL	400	20	-.003	.093	.179	.095	.624
LIMLALL	400	20	-.123	0	.103	.059	.898

[13]Note that the entries for OLS, 2SLSALL and LIMLALL are taken from Angrist, Imbens and Krueger (1995, Table 2). The estimators 2SLSALL and LIMLALL use all $K = 20$ instruments in addition to the constant.

Table 2.2: Case 2, Frequencies

	N=100						N=400					
K	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV
1	4.50	.04	.06	4.00	.06	.06	0	0	0	0	0	0
2	.30	0	0	.12	0	0	0	0	0	0	0	0
3	.04	0	0	.06	.02	0	0	0	0	0	0	0
4	0	0	0	.02	0	0	0	0	0	0	0	0
5	91.30	58.42	73.02	91.64	51.86	69.86	99.34	55.56	73.06	99.06	47.60	71.46
6	2.50	13.52	11.84	2.74	11.90	11.70	.62	13.82	11.32	.86	11.12	10.98
7	.76	8.90	6.24	.84	8.06	5.82	.04	9.34	6.06	.08	7.72	5.26
8	.30	7.34	3.54	.26	5.28	3.84	0	7.66	3.96	0	5.70	3.86
9	.20	5.86	2.98	.08	4.12	2.44	0	6.76	3.24	0	4.30	2.42
10	.10	5.92	2.32	.08	3.58	1.50	0	6.86	2.36	0	3.46	1.50
11				.02	2.66	.98				0	3.42	1.28
12				.04	2.26	.92				0	2.28	.88
13				.04	1.90	.68				0	2.28	.52
14				.02	1.68	.56				0	2.04	.60
15				0	1.38	.44				0	1.84	.32
16				.02	.94	.32				0	1.70	.30
17				0	1.20	.24				0	1.60	.24
18				0	.78	.26				0	1.52	.14
19				.02	1.10	.14				0	1.50	.14
20				0	1.22	.24				0	1.92	.10

Table 3.1: Case 3[14]

			Quantiles				
Est.	N	Max K	.10	.50	.90	MAE	Cov. Rate
OLS	100	*	.113	.166	.223	.166	.012
2SLS	100	10	-.458	.050	.428	.200	.935
LIML	100	10	-.496	.029	.467	.206	.947
Jackknife IV	100	10	-.850	.063	.882	.310	.679
2SLS	100	20	-.460	.082	.429	.192	.897
LIML	100	20	-.522	.050	.514	.200	.935
Jackknife IV	100	20	-.694	.126	.784	.261	.496
2SLSALL	100	20	.033	.147	.268	.148	.474
LIMLALL	100	20	-.572	.097	.784	.253	.869
OLS	400	*	.139	.167	.196	.167	0
2SLS	400	10	-.221	.019	.215	.117	.911
LIML	400	10	-.236	.009	.214	.112	.920
HLIML	400	10	-.249	.007	.218	.115	.943
Jackknife IV	400	10	-.343	-.018	.231	.137	.902
2SLS	400	20	-.247	.038	.211	.118	.889
LIML	400	20	-.259	.015	.219	.113	.927
Jackknife IV	400	20	-.523	-.009	.316	.158	.831
2SLSALL	400	20	.013	.121	.224	.121	.531
LIMLALL	400	20	-.364	.045	.359	.156	.890

[14]Note that the entries for OLS, 2SLSALL and LIMLALL are taken from Angrist, Imbens and Krueger (1995, Table 3). The estimators 2SLSALL and LIMLALL use all $K = 20$ instruments in addition to the constant.

Table 3.2: Case 3, Frequencies

	N=100						N=400					
K	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV	2SLS	LIML	J-IV
1	62.92	59.16	67.62	61.96	52.64	67.16	60.48	64.04	72.20	62.62	56.88	69.74
2	11.18	12.94	11.62	10.02	12.60	11.86	12.72	12.62	11.42	10.58	12.32	11.00
3	6.50	7.38	6.64	6.10	7.72	5.86	7.20	7.52	5.84	6.00	6.94	6.08
4	4.34	4.84	3.80	3.88	4.82	3.60	4.28	4.24	3.84	4.06	5.18	3.56
5	3.86	3.98	2.84	3.40	3.74	2.94	3.90	3.06	2.14	2.90	3.14	2.46
6	2.42	2.84	2.18	2.32	2.80	1.80	2.78	2.26	1.62	2.40	2.28	1.92
7	2.42	2.28	1.46	1.92	2.10	1.52	2.44	1.76	1.00	1.50	1.78	1.08
8	2.06	2.34	1.38	1.32	1.86	1.02	2.16	1.60	.86	1.70	2.00	.94
9	1.92	1.88	1.24	1.22	1.50	.82	2.00	1.48	.72	1.30	1.28	.56
10	2.38	2.36	1.22	.98	1.38	.50	2.04	1.42	.72	1.06	1.10	.50
11				1.02	1.42	.56				.82	.98	.46
12				.80	1.28	.36				.92	.94	.30
13				.86	.84	.38				.44	.68	.30
14				.58	.64	.34				.62	.74	.24
15				.70	.66	.30				.62	.72	.12
16				.76	.60	.30				.42	.62	.10
17				.54	.82	.26				.46	.58	.16
18				.52	.80	.14				.58	.86	..16
19				.48	.80	.12				.36	.36	.20
20				.62	.98	.16				.64	.62	.12

Table 4: Returns to Schooling Estimates[15]

	2SLS		LIML		
Instruments	β	$S(K)$	β	$S(K)$	Cross Val.
4 Regions	.1011	4.1939	.1028	10.1515	10.15157
	(.014)		(.014)		
4 Regions + Qtr-Yr	.0897	4.1954	.0927	10.1514	10.15154
	(.012)		(.013)		
9 Regions	.0920	4.1947	.0956	10.1517	10.15185
	(.013)		(.014)		
9 Regions + Qtr-Yr	.0841	4.1963	.0875	10.1515	10.15166
	(.012)		(.013)		
17 Regions	.0979	4.1963	.1077	10.1515	10.15245
	(.012)		(.014)		
17 Regions + Qtr-Yr	.0892	4.1985	.0973	10.1514	10.15235
	(.012)		(.013)		
32 Regions	.1002	4.2001	.1127	10.1517	10.15300
	(.011)		(.013)		
32 Regions + Qtr-Yr	.0927	4.2033	.1034	10.1515	10.15291
	(.010)		(.012)		
50 Regions	.0985	4.2073	.1143	10.1517	10.15437
	(.010)		(.013)		
50 Regions + Qtr-Yr	.0928	4.2117	.1064	10.1515	10.15428
	(.009)		(.012)		

[15]In this table the entries in the columns with header β are the parameter estimates, with standard errors in parentheses below the estimate. The columns labelled $S(K)$ give the value of the criteria for the set of instruments, while the column headed Cross Val. gives the cross validation statistic from the first stage relationship between number of years of schooling and the instruments.

7076 07

Date Due

OCT 0 9 2008

SEP 3 0 2009

Lib-26-67

MIT LIBRARIES
3 9080 01972 0835